MediatorBot: A Mediator bot for supporting collaborative E-learning using an Intelligent Tutor System

Do Dung Vu
Supervisor: Prof. Sylvie Ratté

Département de génie logiciel et des TI

do-dung.vu.1@ens.etsmtl.ca

April 25, 2019

Overview

Le génie pour l'industrie

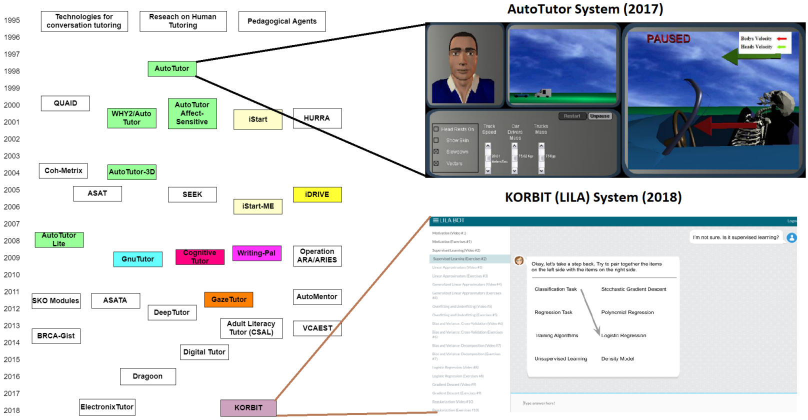
Forecast

Industry Factor	Description
Market size & forecast (Revenue)	USD 402.0 Million (2017) → USD 6,893.4 Million (2024)
Model trend (2017)	Learner Model – 65.33% Pedagogical Model – 22.76% Domain Model – %
Deployment trend (2017)	On-Premise – 84.63% Cloud – 15.37%
Technology trend (2017)	Machine Learning – 22.00% Deep Learning – 5.07% NLP – 68.28% Others – 4.65%
Application trend (2017)	Learning Platform & Virtual Facilitators– 56.37% Intelligent Tutoring System – 21.73% Smart Content – 15.06% Fraud & Risk Management – 2.69% Others – 4.15%
End-Use trend (2017)	Higher Education – 52.37% K-12 Education – 33.98% Corporate Training – 13.66%
Regional trend (2017)	North America – 60.16% Europe – 18.62% Asia Pacific – 16.48% LA – 1.22% MEA– 3.52%

AI in Education industry 3600 synopsis, 2013 – 2024

Source: AAAI, IEEE, WEF, IAAIL, Company Annual Reports, Hoovers, Primary Interviews, Global Market Insights

History



Source: (1) autotutor.org (2) lilabot.com

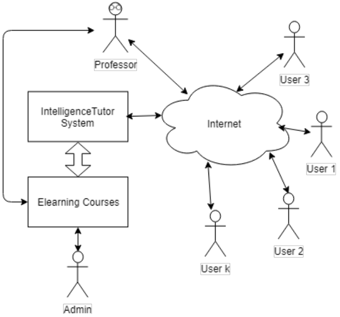
ITS

ITS main purposes:

- Help students construct expressions of material as answers to questions and solutions to solve the challenging problems
- Ask questions that tap deep levels of reasoning and that involve collaboration
- Solve problems that involve deep argumentation

Context

- Online group learning on a given domain-specific (e.g., statistic)
- The group must discuss about a given topic or assignment [*]
- The Intelligence Tutor System (ITS) helps Professor to monitor the progress of students and Admin to encourage their study



[*]https://mydalite.org/en/

Literature survey

- An ITS monitors students' knowledge, skills, and psychological characteristics and response [1] - 2014
- Conversational agents have talking heads that speak, point, gesture, and exhibit facial expressions. [2] - 2016

[1] Sottilare, R, Graesser, AC, Hu, X, Goldberg, B (Eds.) (2014). Design recommendations for intelligent tutoring systems: instructional management, (vol. 2). Orlando: Army Research Laboratory

[2] Johnson, WL, & Lester, JC. (2016). Face-to-face interaction with pedagogical agents, twenty years later. International Journal of Artificial Intelligence in Education, 26(1), 25–36.

Literature survey

- AutoTutor and its progenies [3] help students learn by holding a conversation in natural language - 2016
- Agent intervention aiming to link students' contributions to previously acquired knowledge can improve both individual and group studying when implemented in the context of a collaborative learning activity in higher education [4]- 2017

[3] Graesser, AC. (2016). Conversations with AutoTutor help students learn. International Journal of Artificial Intelligence in Education, 26, 124–132

[4] Tegos, S., & Demetriadis, S. (2017). Conversational Agents Improve Peer Learning through Building on Prior Knowledge.Educational Technology & Society, 20(1), 99–111

Existing problem

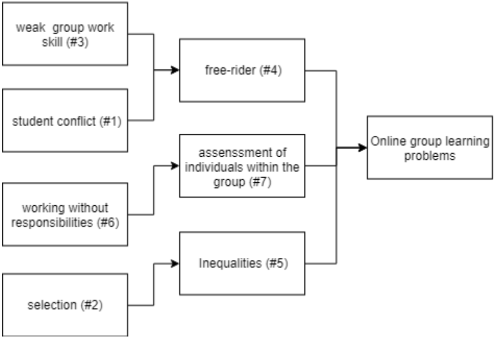
The seven most commonly in Online Learning found in the literature are the following [5]

- (1): the student has conflicts works in the group
- (2): the selection of the groups is not good
- (3): the students don't have enough group-work skills
- (4): some students want to work alone or become the free-riders
- (5): the possible inequalities of student abilities appears in the group
- (6): some members do not commit to working in the group with their responsibilities
- (7): the assessment of individuals within the groups is not fair

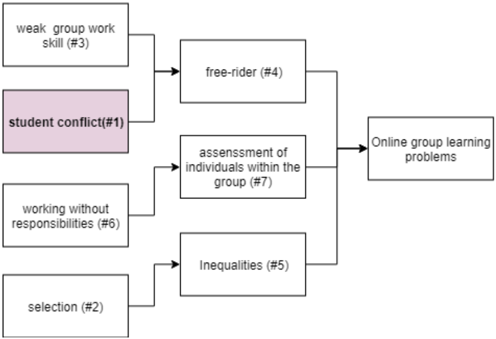
[5] Jianxia Du, Chuang Wang, Mingming Zhou, Jianzhong Xu, Xitao Fan & Saosan Lei (2018) Group trust, communication media, and interactivity: toward an integrated model of online collaborative learning, Interactive Learning Environments, 26:2, 273-286,

Interaction of problems

Most of these problems above of online group learning are inter-related



Main problem



#3: solved by orientation training from admin
#6, #2: solved by professor
#1: solved by ITS system

→ **We want to solve the problem of student conflicts by using the ITS.**

Scenario

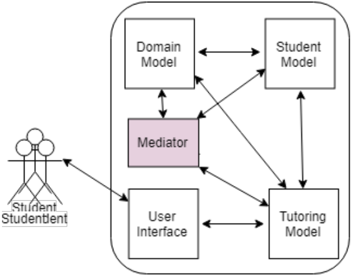


Figure: ITS with Mediator

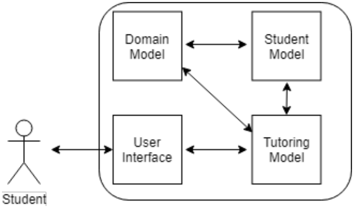
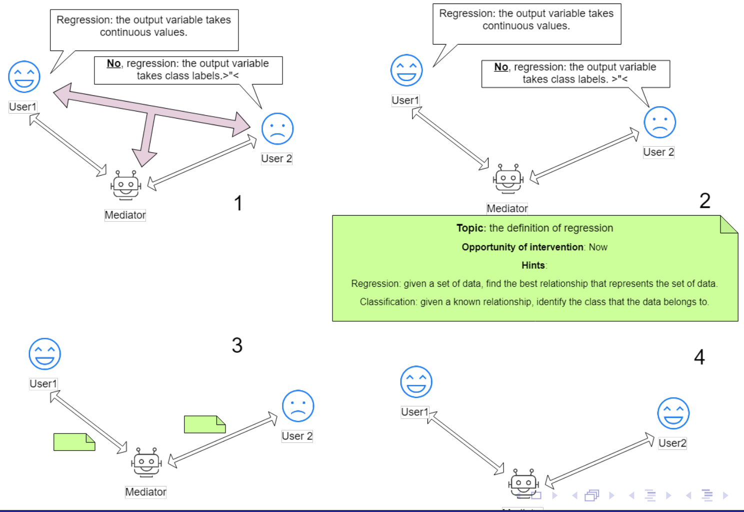


Figure: Original ITS [6]

[6] N. T-Nghe and L. S-Thieme, "Multi-Relational Factorization Models for Student Modeling in Intelligent Tutoring Systems", 17th International Conference on Knowledge and Systems Engineering (KSE) 2015

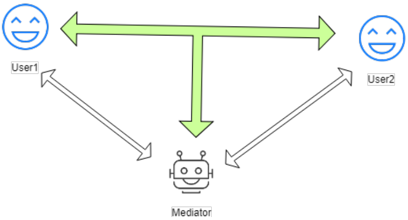
Intuitive Scenario



MediatorBot

MediatorBot identifies the debated problem, the opportunities for intervention, generates the hints, and answers the related topic question of students to encourage the users to collaborate more effectively in the online group learning with low price in the specific-domain

Motivations



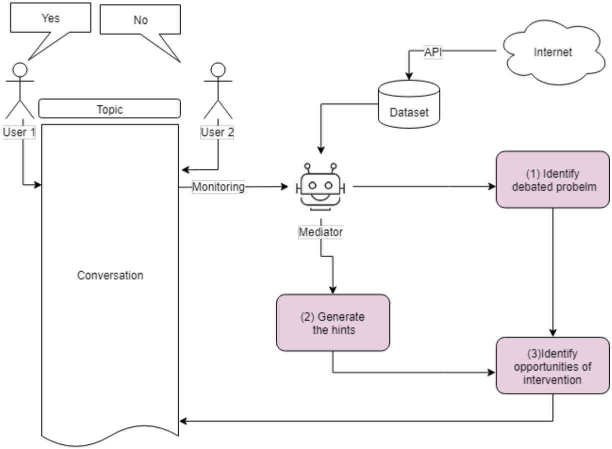
- Future state-of-the-art interventions with low price for group online learning
- Encourage student collaboration online
- Easily scalable

Objectives

Main objective: Propose a smart Mediator to support constructive discussion based on the Intelligent Tutor System:

- Generate hints to help users solve the topic or problem automatically
- Identify the debated problem
- Intervene in the conversation to resolve the conflict

Mediator system

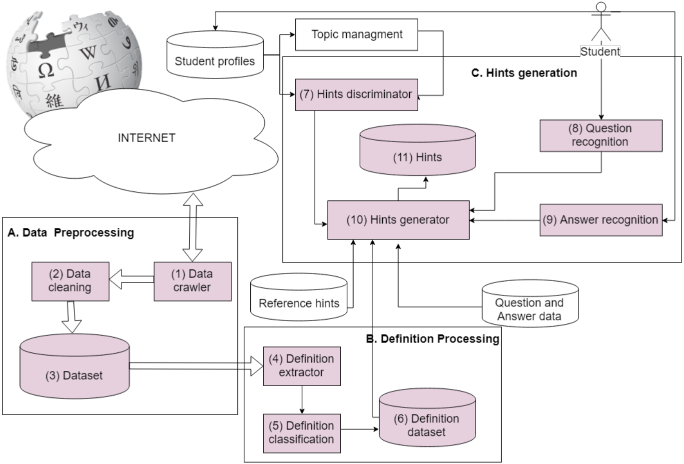


(1) Generate hints to help users solve the topic or problem automatically

(2) Identify the debated problem

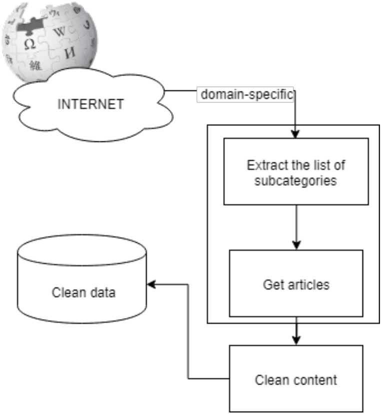
(3) Intervene in the conversation to clarify the problem

Objective 1— Structure



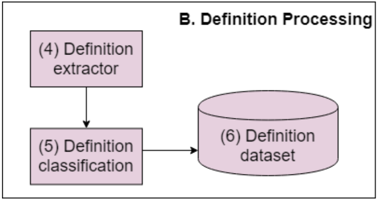
Objective 1 — Methodology — A. Data preprocessing

(1) Data crawler: crawling data from wikipedia with a given domain-specific (e.g., statistic)
(2) Data cleaning: clean the unicode, convert xml equation to latex equation, clean punctuation, split raw text to line by line sentence
(3) Dataset: save data to the tsv file with it fields: title, link, content

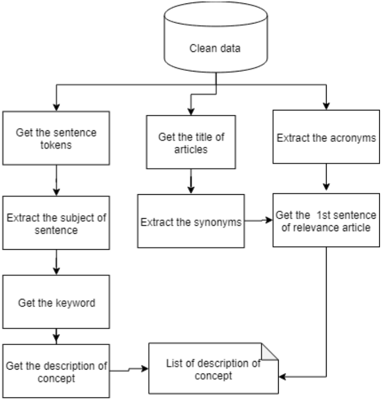


Objective 1 — Methodology — B. Definition processing

(4) Definition extractor: Extract the description of each concept
(5) Definition classification: classify type of definition based on supervisor algorithm (Positive/Negative)
→ Using the oversampling methodology to reweight the Positive and Negative samples
(6) Definition dataset



Objective 1 — Methodology — B. Definition processing — Definition extractor



- Split raw text dataset to the sentence tokens
- Extract the technical keyword acronyms and synonyms
- Extract subject (noun phase) $\rightarrow$ Get the keyword (concept)
- Get the description of concept
- Save the list (dict) description of concept which is called dictionary of definition

Objective 1 — Methodology — B. Definition processing — Definition extractor

* Extract the acronyms: considering the algorithm proposed in [7]
* Extract the synonyms: using the Wikipedia API of direction searching [8]
* Extract the subjects: using the industrial-strength natural language processing frame work [9]

[7] A. Schwartz and M. Hearst (2003) A Simple Algorithm for Identifying Abbreviations Definitions in Biomedical Text. Biocomputing, 451-462

[8] http://wikisynonyms.ipeirotis.com/api

[9] https://spacy.io/

Objective 1 — Methodology — B. Definition processing — Definition extractor — Examples

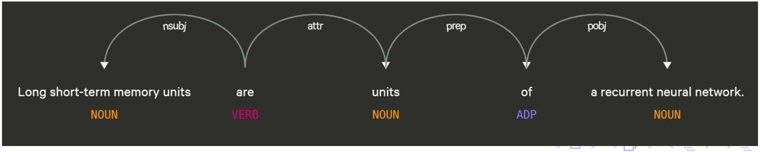
* **Synonyms:** Linear regression model

'Linear modeling','Regression coefficient','Linear Regression', 'Regression coefficients','Regression line','Linear weights','Multiple linear regression','Line regression','Linear regression model','Linear trend','Multi-linear regression','Linear fit','Line of regression','Linear regression'

* **Acronyms:** RNN

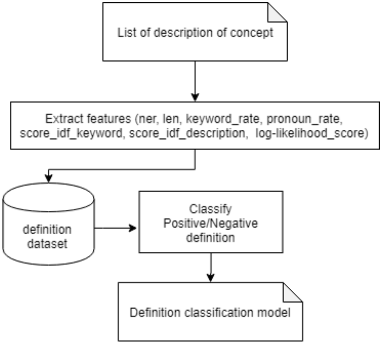
'recurrent neural network','random neural network'

* **Tree of sentence:**



Objective 1 — Methodology — B. Definition processing — Definition classification

- Extract the features → score table
- Save the score table to the definition dataset
- Classify the Positive/Negative definition based on logistic regression
- Save the classification model



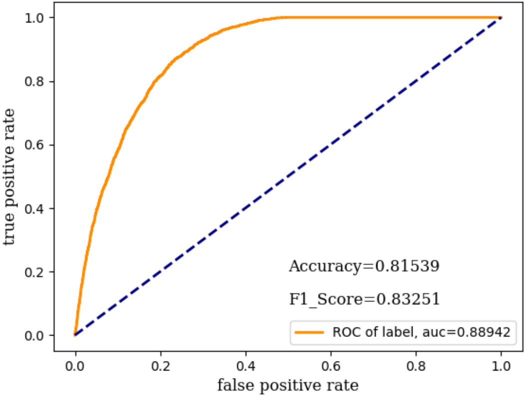
→ **Create the resources for Hints generator**

Objective 1 — Methodology — B. Definition processing — Definition classification

Features	Summary
`length_of_keyword`	the number words in the keyword
`length_of_description`	the number words in the description
`score_keyword`	inverse document frequency of keyword
`score_description`	inverse document frequency of concepts description
`ner_in_description`	name of entity recognition within the description
`coreference_in_description`	compute the coreference resolution score
`type_of_word`	recognize type of word (verb, noun, etc.,)
`non_of_word`	recognize the none of word (symbol, number, etc.,)
`pronouns_rate`	the rate of $\frac{pronouns}{nouns}$
`keyword_rate`	the rate of $\frac{keyword\_position}{length\_of\_description}$
`perplexity`	the real value of perplexity of desciption
`likelihood_score_description`	the log-likelihood probability score of description based on sum of probability term by using language model based on RNN

Table: Features of definition

Objective 1 — Methodology — B. Definition processing — Definition classification— Performance

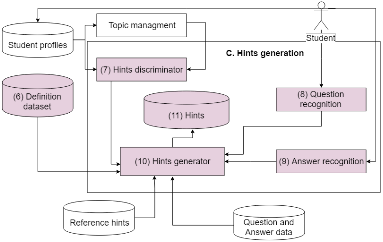


Objective 1 — Methodology — B. Definition processing — Definition classification— Example

key	definitnion	label	score keyword	score description	...	likelihood score description
Linear regression	In statistics, linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).	**Positive**	14.58	130.53	...	10.64
linear regression models	The numerical methods for linear least squares are important because linear regression models are among the most important types of model, both as formal statistical models and exploration of data-sets.	**Negative**	20.78	146.77	...	10.49
local linear regressions	local linear regressions are preferred because they have better bias properties and have better convergence.	**Negative**	25.93	85.57	...	8.04

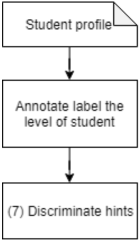
Objective 1 — Methodology — C. Hint generation

(7) Hints discriminator: classify level of hints based on the student profiles
(8) Question recognition: recognize question of student
(9) Answer recognition: recognize answer of student
(10) Hints generator: generate hint based on hint types, level, and language model



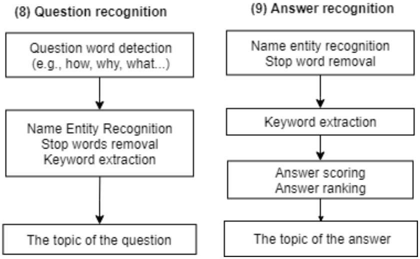
Objective 1 — Methodology — C. Hint generation — Hints discriminator

- Classify the students' level based on their profile
- Discriminate hints based on level of student

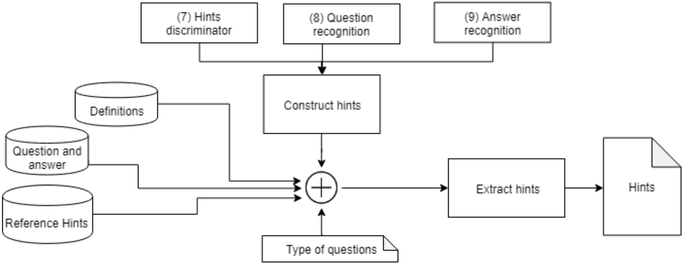


Objective 1 — Methodology — Hint generating — Question & Answer recognition

(8) Question recognition: Recognize the users' questions
(9) Answer recognition: Recognize the users' answers



Objective 1 — Methodology — C. Hint generation— Hint generator



Objective 1 — Methodology — C. Hint generation— Hint generator— Construct hints

* Type of questions: definition, definition & explanation, contrast, explanation, and deeper understanding
* The hints are phrased in the form of "Think about X" or "Consider X" where X is the part of expectation answer.
* Using Linear regression model based on the features:

Features	Summary
length_of_hint	the number words in the hint
overlap_question_hint	the rate of overlap between question and hint
score_keyterm	inverse document frequency of keyterm in hint
keyhint_keyquestion_ratio	the ratio of $\frac{number\_of\_keyhint}{number\_of\_keyquestion}$
topic_overlap	content overlap between the question and hint
pronouns_rate	the rate of $\frac{pronouns}{nouns}$ in hint
keyword_rate	the rate of $\frac{keyword\_position}{length\_of\_hint}$
perplexity	the real value of perplexity of hint
ner_in_hint	name of entity recognition within the hint
score_of_hint	the log-likelihood probability score of hints based on sum of probability terms by using language model based on RNN

Table: Features of hints

Objective 1 — Methodology — C. Hint generation— Hint generator— Construct hints— Example

Question:

You are given a dataset of images of wildlife in Africa.
You are tasked with building a model which can identify animals in the images.
Is this a regression or classification problem? Explain why?

Answer:

It is the regression problem because the animal is the independent entity in the africa

Hints:

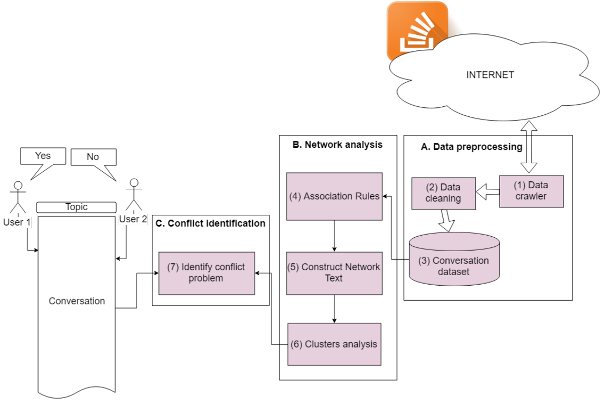
- Recall that each animal is a class.
- Recall that each animal is a discrete class.
- Consider that each animal is a separate class.
- Consider that we are choosing between a set of categories.
- Think about the following: we are choosing between discrete-valued output variables.
- Consider that each image can contain several animals, and therefore the model must predict the existence of each type of animal.

(1) Generate hints to help users solve the topic or problem automatically

(2) Identify the debated problem

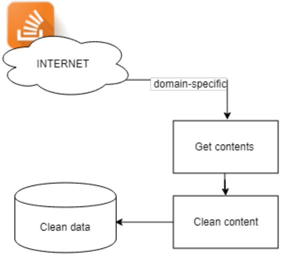
(3) Intervene in the conversation to clarify the problem

Objective 2 — Structure



Objective 2 — Methodology — Data preprocessing

(1) Data crawler: crawling data from stackoverflow, hangout, messenger, slack with a given domain (e.g., statistic)
(2) Data cleaning: Clean content: clean unicode, equation over the conversation
(3) Conversation dataset: save the conversation dataset



Objective 2 — Methodology

(4) Association rules [10]: find the interesting association or correlation relationship between dominant words [10]

$$Rule: X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

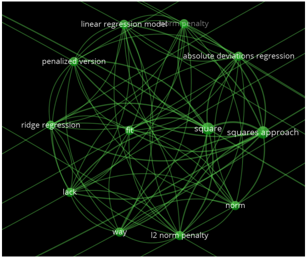
$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

- *Support*: how frequently the itemset appears in the dataset.
- *Confidence*: how often the rule has been found to be true.
- *Lift*: the ratio of the observed support to that expected if X and Y were independent

[10] A. Alamsyah, M. Paryasto, F. J. Putra, R. Himmawan, "Network text analysis to summarize online converstations for marketing intelligence efforts in telecommunication industry", in 2016 ICoICT

Objective 2 — Methodology



An example of text network

(5) Construct network text of dominant word: include weighted edge result for association rule processes

(6) Network analysis: create context, keyword, and sense from network text → employ centrality to find the most influential words in the networks and modularity to find words cluster/ groups in the network

Objective 2 — Methodology

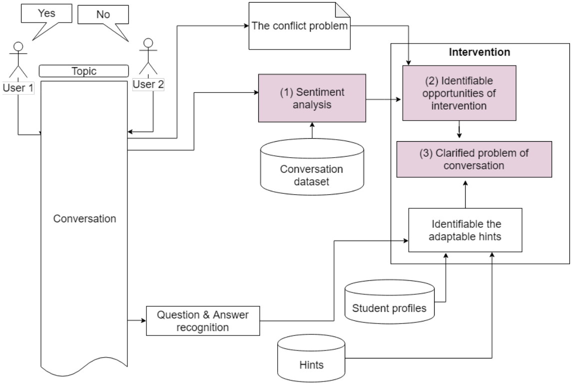
+ Using K-means algorithm to cluster and store Association Rules without favoring or excluding any evaluation measurement of an association rule.[2]

→**(7) Identify conflict problem:** get the conflict problem related to the topic by mapping conversation to clusters analysis

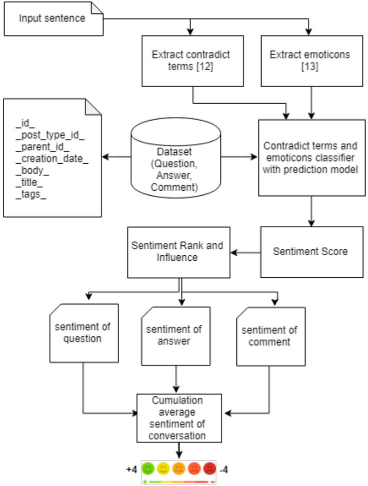
[11] A. Dahbi, M. Mouhir, Y. Balouki and T. Gadi,"Classification of association rules based on K-means algorithm", in 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)

(1) Generate hints to help users solve the topic or problem automatically
(2) Identify the debated problem
(3) **Intervene in the conversation to clarify the problem**

Objective 3— Structure



Objective 3 — Methodology — Sentiment analysis



- Listening the conversation
- Using SVM in classifying the Emoticons of content
- Cumulate the sentiment of question, answer, and comment for evaluating the sentiment of conversation

[12] M. Marneffe, A. N. Rafferty, and C. D. Manning. 2008. Finding contradictions in text. In Proc. ACL

[13] L. Ling, S. Larsen, "Sentiment Analysis on Stack Overflow with Respect to Document Type and Programming Language", KTH ROYAL INSTITUTE OF TECHNOLOGY, 2018

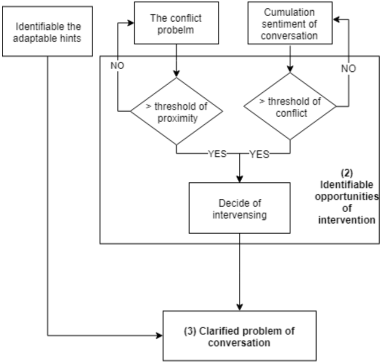
Objective 3 — Methodology — Sentiment analysis — Preprocessing

1. Remove the bodies of all $< pre >, < code >$ and $< blockquote >$ tags.
2. Extract all text (excluding any HTML tags).
3. Replace every sequence of consecutive whitespace (such as spaces, tabs and line feeds) with a single space character each.
4. Remove all http and https urls.
5. Strip leading and trailing whitespace
6. Use the Senti4SD [14] to classify the dataset

[14] Calefato, F., Lanubile, F., Maiorano, F., Novielli N. (2018) "Sentiment Polarity Detection for Software Development," Empirical Software Engineering, 23(3), pp:1352-1382

Objective 3 — Methodology — Intervention

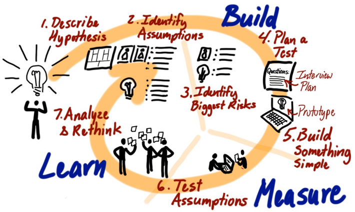
(2) Identifiable opportunities of intervention: analysis the serious of conversation and conflict problem
(3) Clarified problem of conversation: give the right intervention



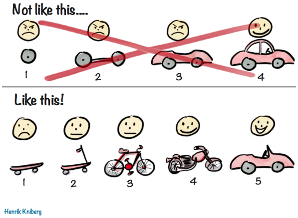
Evaluation measurement

Because this is the conversation between Human and machine, so we prefer to use the users' experiment test to get feedback score in range (1,5) and expert recommendations.

Evaluation — Approach



Source: https://www.jpattonassociates.com/



Source: https://quickleft.com

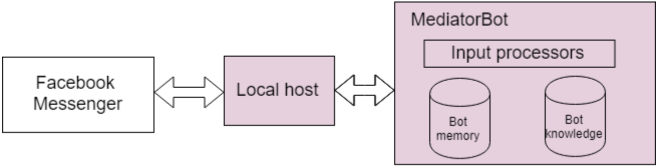
→ We evaluate our system by using the user experiments testing.

- students' experiments
- professor recommendations

→ Users: students at the class offline, students on LILA[4], friends (if REB is valid) or Amazon Mechanical Turk[5]

4.https://lilabot.com 5.https://www.mturk.com/

Evaluation — Environment



- (1) Use the Facebook messenger API[1] to set up the conversation environment
- (2) Set up the flask server for local host
- (3) Process the conversation with the given bot memory and knowledge

[1] https://developers.facebook.com

Evaluation — Environment

- (4) Feedback statistic evaluation (`https://docs.gogle.com/forms/u/0/`)
- (5) Using Cohen's Kappa [15] for evaluating the agreement of human and machine experiment

[15] J. Toppi and N. Sciaraffa and Y. Antonacci and A. Anzolin and S. Caschera and M. Petti and D. Mattia and L. Astolfi, "Measuring the agreement between brain connectivity networks", in 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)

Achievements

(1) 3 years Mitacs accelerate grant for Natural Language Generation for Intelligent Tutoring Systems

(2) Directly apply the results to LILA and Korbit pipeline at Ai-educate Inc
`https://lilabot.com/`

Work plan

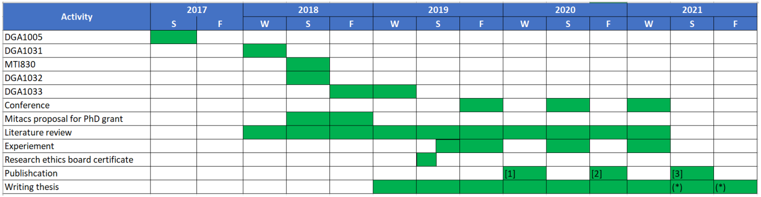


Figure: Work schedule

Journals:

[1] Journal of Artificial Intelligence Research

[2] Technology, Knowledge and Learning

[3] Education and Information Technologies

Finished courses:

(1) DGA1005

(2) MTI830

Thank You

Appendices

ASAT: AutoTutor Script Authoring Tool is the primary authoring tool for AutoTutor. Can direct multiple agents and external events/controls.
ASATA: AutoTutor Script Authoring Tool for Assessment is a specialized authoring tool developed with the Educational Testing Service for developing for building dialog-based high stakes assessments.
AutoMentor (STEM Thinking): Uses epistemic analysis of discourse in student group chats to help students learn how to think and act like STEM (science, technology, engineering, and mathematics) professionals in a multi-party serious game simulation of urban planning.
AutoTutor (Computer Literacy): Core AutoTutor natural language tutoring system, which uses expectation-misconception dialog and deep questions, latent semantic analysis & regular expressions, and talks with user through the animated agent(s).
AutoTutor-3D (Physics): An extension of AutoTutor for physics, AutoTutor-3D added interactive three dimensional simulations of physics problems designed in 3D Studio Max.
AutoTutor Affect-Sensitive (Computer Literacy): AutoTutor-AS detected affect using natural language and discourse, facial expressions, body posture, and speech. Feedback considered student emotions and cognitive states. Sometimes called AutoTutor-ES (Emotion Sensitive).
AutoTutor Lite (General): AutoTutor Lite (ATL) is a web-based variant of AutoTutor designed for simpler authoring, rapid deployment, and integration into thirdparty systems.
BRCA-Gist (Breast Cancer Risk): An AutoTutor Lite tutor led by the Miami University, intended to tutor understanding of risk probabilities and personal breast cancer risk.
Coh-Metrix: A linguistic analysis toolkit with over 200 metrics. The "Coh" stands for cohesion and coherence.
CSAL Adult Literacy Tutor (Reading): This tutoring system project for the Center for the Study of Adult Literacy (CSAL) is intended to help learners who 460 Int J Artif Intell Educ (2014) 24:427–469 struggle with print media, through closer integration of trialogs, web pages, and multimedia.
DeepTutor (Physics): Tutor that uses learning progressions to foster deep learning of physics concepts, as well as enhanced semantic analysis, such as entailment.

Appendices

GazeTutor (Biology): Enhanced version of Guru Tutor that monitors and reacts to student gaze.
Gnu Tutor (General): An open source Java release of an early version AutoTutor Lite.
Guru Tutor (Biology): Tutoring system for biology designed based on observation of expert tutors. Uses collaborative lecturing and concept maps to support learning.
HURAA (Research Ethics): The Human Use Regulatory Affairs Advisor for training ethics in human experiments. AutoTutor agents helped navigate hypertext multimedia containing case-based reasoning and multiple information retrieval mechanisms.
iDRIVE (Computer Literacy, Physics, Biology): Instruction with Deep-Level Reasoning Questions in Vicarious Environments where the learner observes two pedagogical agents demonstrate deep explanations and model effective learning behavior (e.g. question-asking).
iSTART (Reading): Interactive Strategy Training for Active Reading and Thinking is a tutoring system for improving reading comprehension by training reading strategies. Uses multi-agent conversations and specialized semantic analysis to tutor reading strategies.
iSTART-ME (Reading): The Motivationally-Enhanced (ME) version of iSTART provides tutoring using an interactive game environment.
MetaTutor (Biology): Tutors self-regulated learning (SRL) skills inside a hypermedia setting.
Operation ARA (Scientific Reasoning): Operation Acquiring Research Acumen is an extension of the Operation ARIES project that adds additional features and game content.
Operation ARIES (Scientific Reasoning): Operation Acquiring Research, Investigative, and Evaluative Skills is a trialog-based tutoring system and serious game for teaching critical thinking. Learners resolve inconsistent information about scientific methods inside a serious game narrative.

Appendices

QUAID: Question Understanding Aid was a tool to evaluate the comprehensibility of questions.
SEEK Web Tutor (Critical Thinking): The Source, Evidence, Explanation, and Knowledge Tutor was designed to help learners evaluate the credibility and relevance of information using tutoring-enhanced web search, with spoken hints, pop-up ratings and metacognitive journaling.
SKO Modules (General): Sharable Knowledge Object Modules are encapsulated, cloud-hosted modules that compose web services to provide tutoring. Currently being applied to Algebra.
VCAEST (Medical): Virtual Civilian Aeromedical Evacuation Sustainment Training is designed to train civilian medical personnel on federal guidelines for emergency situations and triage.
WHY2/AutoTutor (Physics): Extension of AutoTutor that approached tutoring conceptual physics. This was part of a larger WHY2 project that included Int J Artif Intell Educ (2014) 24:427–469 461
WHY2/Atlas. WHY2 was a reference to an old tutoring system called WHY and the year 2000 (e.g., Y2K).
Writing-Pal (Writing): This tutor attempts to improve essay and academic writing

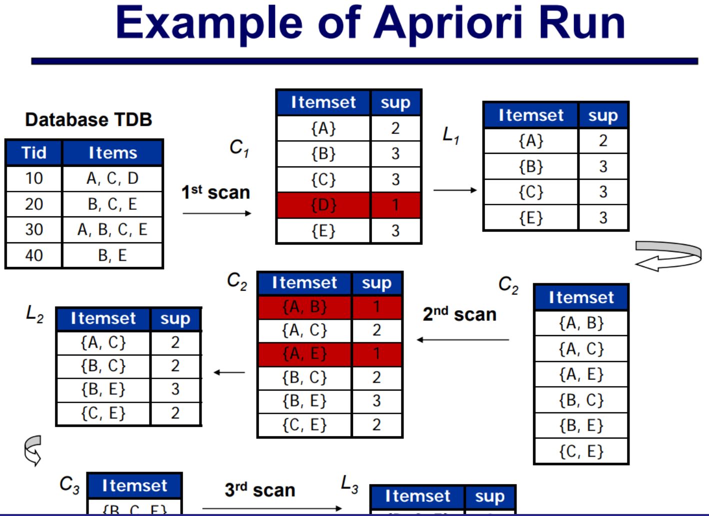
skills and provides automated evaluation and feedback on essays. It is related to the iSTART system.

Appendices

TID	s	c
bread ⇒ peanut-butter	0.60	0.75
peanut-butter ⇒ bread	0.60	1.00
beer ⇒ bread	0.20	0.50
peanut-butter ⇒ jelly	0.20	0.33
jelly ⇒ peanut-butter	0.20	1.00
jelly ⇒ milk	0.00	0.00

TID	Items
T1	bread, jelly, peanut-butter
T2	bread, peanut-butter
T3	bread, milk, peanut-butter
T4	beer, bread
T5	beer, milk

Appendices


Example of Apriori Run
Database TDB
Tid | Items
10 | A, C, D
20 | B, C, E
30 | A, B, C, E
40 | B, E
1st scan
C1
Itemset | sup
{A} | 2
{B} | 3
{C} | 3
{D} | 1
{E} | 3
L1
Itemset | sup
{A} | 2
{B} | 3
{C} | 3
{E} | 3
C2
Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}
2nd scan
C2
Itemset | sup
{A, B} | 1
{A, C} | 2
{A, E} | 1
{B, C} | 2
{B, E} | 3
{C, E} | 2
L2
Itemset | sup
{A, C} | 2
{B, C} | 2
{B, E} | 3
{C, E} | 2
C3
Itemset
3rd scan
L3
Itemset | sup

* We extract contracdiction features on which we apply logistic
Features for constradiction detection

Features	Summary
`polarity`	The polarity features capture the presence (or absence) of linguistic markers of negative polarity contexts
`numeric`	The numeric features recognize (mis-)matches between numbers, dates, and times
`antonymy`	list of antonyms and contrasting words comes from WordNet, from which we extract words with direct antonymy links and expand the list by adding words from the same synset as the antonyms
`structural`	determine whether the syntactic structures of the text and hypothesis create contradictory statements.
`factivity`	The context in which a verb phrase is embedded may give rise to contradiction
`modality`	Simple patterns of modal reasoning are captured by mapping the text and hypothesis to one of six modalities ((not)possible, (not)actual, (not)necessary), according to the presence of predefined modality markers such as can or maybe

Table: Features of contradict detection

Appendices

Emoji	Name(s)	Score	Emoji	Name(s)	Score	Emoji	Name(s)	Score	Emoji	Name(s)	Score	Emoji	Name(s)	Score	Emoji	Name(s)	Score
	imp	-4		face_with_head_bandage	-2		ghost	-1		triumph	0		kissing_smiling_eyes	2		heart	3
	middle_finger; fu	-4		fearful	-2		hushed	-1		upside_down_face	0		lips	2		heart_eyes	3
	pouting_cat	-4		flushed	-2		mask	-1		handshake	1		ok_hand	2		heart_eyes_cat	3
	rage; pout	-4		frowning_face	-2		nerd_face	-1		laughing; satisfied	1		relaxed	2		heartbeat	3
	angry	-3		grimacing	-2		pensive	-1		pray	1		relieved	2		heartpulse	3
	anguished	-3		lying_face	-2		roll_eyes	-1		slightly_smiling_face	1		smile	2		innocent	3
	broken_heart	-3		nauseated_face	-2		slightly_frowning_face	-1		stuck_out_tongue	1		smile_cat	2		joy	3
	hankey; poop; shit	-3		open_mouth	-2		stuck_out_tongue_winking_eye	-1		sunglasses	1		smiley	2		joy_cat	3
	scream	-3		persevere	-2		sweat	-1		+1; thumbsup	2		smiley_cat	2		kissing_heart	3
	scream_cat	-3		skull	-2		thinking	-1		astonished	2		smirk	2		purple_heart	3
	smiling_imp	-3		skull_and_crossbones	-2		zipper_mouth_face	-1		blush	2		smirk_cat	2		revolving_hearts	3
	sob	-3		sneezing_face	-2		clown_face	0		cowboy_hat_face	2		sweat_smile	2		sparkling_heart	3
	worried	-3		tired_face	-2		drooling_face	0		crossed_fingers	2		v	2		two_hearts	3
	-1; thumbsdown	-2		unamused	-2		expressionless	0		grin	2		100	3		wink	3
	cold_sweat	-2		weary	-2		money_mouth_face	0		grinning	2		black_heart	3		yellow_heart	3
	confounded	-2		disappointed_relieved	-1		neutral_face	0		hugs	2		blue_heart	3		yum	3
	confused	-2		dizzy_face	-1		no_mouth	0		kiss	2		clap	3		raised_hands	4
	cry	-2		face_with_thermometer	-1		sleeping	0		kissing	2		cupid	3		rofl	4
	crying_cat_face	-2		fist_oncoming; facepunch; punch	-1		sleepy	0		kissing_cat	2		gift_heart	3			
	disappointed	-2		frowning	-1		stuck_out_tongue_closed_eyes	0		kissing_closed_eyes	2		green_heart	3			

Table of Emoticon

Advantages of Logistic Regression:
- Lots of ways to regularize your model, and you don't have to worry as much about your features being correlated, like you do in Naive Bayes.
- You also have a nice probabilistic interpretation, unlike decision trees or SVMs, and you can easily update your model to take in new data (using an online gradient descent method), again unlike decision trees or SVMs.
- Use it if you want a probabilistic framework (e.g., to easily adjust classification thresholds, to say when you're unsure, or to get confidence intervals) or if you expect to receive more training data in the future that you want to be able to quickly incorporate into your model.

Algorithm	Problem Type	Results interpretable by you?	Easy to explain algorithm to others?	Average predictive accuracy	Training speed	Prediction speed	Amount of parameter tuning needed (excluding feature selection)	Performs well with small number of observations?	Handles lots of irrelevant features well (separates signal from noise)?	Automatically learns feature interactions?	Gives calibrated probabilities of class membership?	Parametric?	Features might need scaling?
KNN	Either	Yes	Yes	Lower	Fast	Depends on n	Minimal	No	No	No	Yes	No	Yes
Linear regression	Regression	Yes	Yes	Lower	Fast	Fast	None (excluding regularization)	Yes	No	No	N/A	Yes	No (unless regularized)
Logistic regression	**Classification**	**Somewhat**	**Somewhat**	**Lower**	**Fast**	**Fast**	**None (excluding regularization)**	**Yes**	**No**	**No**	**Yes**	**Yes**	**No (unless regularized)**
Naive Bayes	Classification	Somewhat	Somewhat	Lower	Fast (excluding feature extraction)	Fast	Some for feature extraction	Yes	Yes	No	No	Yes	No
Decision trees	Either	Somewhat	Somewhat	Lower	Fast	Fast	Some	No	No	Yes	Possibly	No	No
Random Forests	Either	A little	No	Higher	Slow	Moderate	Some	No	Yes (unless noise ratio is very	Yes	Possibly	No	No
AdaBoost	Either	A little	No	Higher	Slow	Fast	Some	No	Yes	Yes	Possibly	No	No
Neural networks	Either	No	No	Higher	Slow	Fast	Lots	No	Yes	Yes	Possibly	No	Yes

Table of supervisor learning algorithm comparison

Appendices

Using wikiSynonyms API to get the synonyms keywords

E.g., Linear regression model 'Linear modeling', 'Regression coefficient', 'Linear Regression', 'Regression coefficients','Regression line', 'Linear weights', 'Multiple linear regression', 'Line regression', 'Linear regression model', 'Linear trend', 'Multi-linear regression', 'Linear fit', 'Line of regression', 'Linear regression'

Appendices

TABLE 1 : SOME INTERESTINGNESS MEASURES

Measures	Formula
Lift(Lift)	$Lift(X \to Y) = \frac{P(XY)}{P(X)P(Y)}$
Information Gain(GI)	$GI(X \to Y) = \log_2 \frac{P(XY)}{P(X)P(Y)}$
Example & Counter Example Rate(ECR)	$ECR(X \to Y) = 2 - \frac{1}{conf(X \to Y)}$
Jaccard(JRD)	$JRD(X \to Y) = \frac{P(XY)}{P(X\overline{Y}) + P(Y)}$
Cosinus(COS)	$COS(X \to Y) = \frac{P(XY)}{\sqrt{P(X)P(Y)}}$
Pearl(PRL)	$PRL(X \to Y) = P(X)\mid P(\frac{Y}{X}) - P(Y)\mid$
Loevinger(LVG)	$LVG(X \to Y) = \frac{P(\frac{Y}{X}) - P(Y)}{1 - P(Y)}$
Conviction(CNV)	$CNV(X \to Y) = \frac{P(X)P(\overline{Y})}{P(XY)}$
Zhang(ZHN)	$ZHN(X \to Y) = \frac{P(XY) - P(X)P(Y)}{\max\{P(XY)P(\overline{Y}), P(Y)P(X\overline{Y})\}}$
Piatetsky Shapiro(PS)	$PS(X \to Y) = P(XY) - P(X)P(Y)$
Sebag-Schoenauer(SBG)	$SBG(X \to Y) = \frac{P(XY)}{P(X\overline{Y})}$