NLP Applications in Education

LING 575 Project 1 Presentation, Weifeng Jin

Jan 30 2018

Education

- Definition from Wikipedia: Education is the process of facilitating learning, or the acquisition of knowledge, skills, values, beliefs, and habits.

- Recent development on technology, including NLP techniques, plays an significant role in the field of educational research, including evaluation, learning methods development and feedback analysis.

Education Research

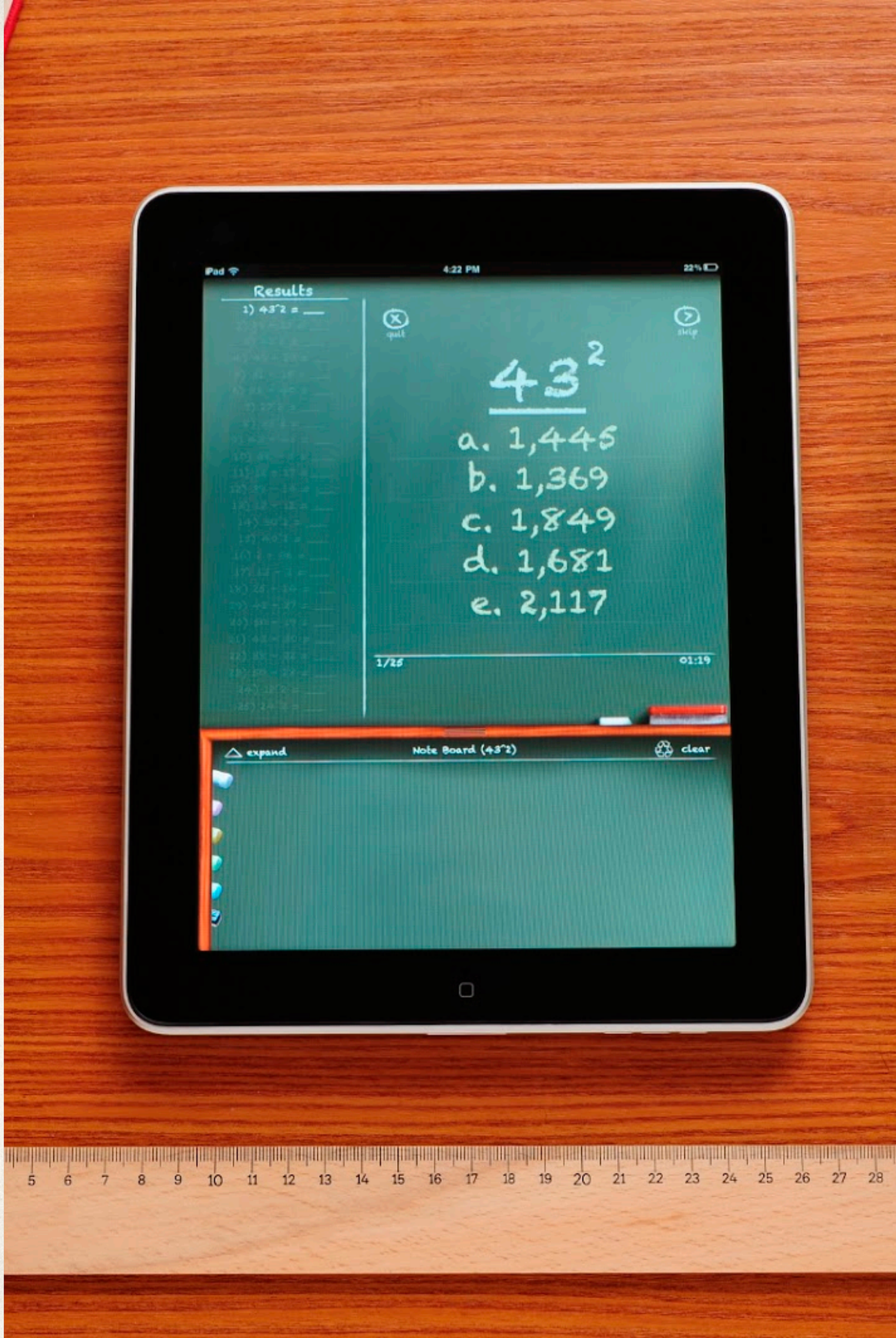
- Definition from American Educational Research Association: education research is the scientific field of study that examines education and learning processes and the human attributes, interactions, organizations, and institutions that shape educational outcomes.

- Qualitative Research: narrative/historical research

- Quantitative Research: experimental/descriptive survey/correlational research

Recent Development in Education

- Modes of technology-based teaching
- Computer-mediated communication
- More advanced classroom tools
- Commercial administrative systems



NLP Applications in Education

- **Plagiarism Detection**: improve the fairness of educational evaluation and protect the academic integrity using NLP techniques

- **Automatic Essay Grading**: improve the accuracy of essay evaluation and save manpower

- **Automatic Generation of Problems/Poems**: a useful application of artificial intelligence to improve teaching and boost creativity

Related Conferences and Workshops

- International Educational Technology Conference (IETC)
- Future of Education Technology (FETC)
- International Society for Technology in Education (ISTE)

Shared Tasks in Education Domain

- Standard evaluation test collection
- Textbooks and curriculum development
- Instructor training and evaluation

Differences between NLP and Education Domain

Research Subjects

- For NLP research, the subjects are generally natural languages and their attributes

- For education research, we are developing a better methodology, an evaluation system, or textbooks for human beings

- The interaction between researchers and subjects are more important in the field of education research

Domain Scope

- Education has a far larger scope than NLP research. It involves economics, politics, system design, more advanced electrical systems and many different areas.

- NLP cannot solve all problems in the education domain.

- NLP is just a tool to help researchers understand some of the problems related with natural languages in the domain of education.

- **Problems in education that can be addressed by NLP**: use technology to save human power, develop automatic systems for evaluation, supplement more high-quality instructive material to the textbook

- **Problems in education that cannot be addressed by NLP**: how to develop an effective and competitive national education system, how can universities find more funding resources, how can schools create an equal environment for learning and teaching

Challenges in the Education Domain

Controlling Costs

- Imbalance in resource allocation for different schools
- Most schools: funding shortage, unattractive for great instructors and students
- Best schools: fund research and teaching activities, maintain credibility and reputation
- => Use NLP techniques to save resources and control costs, especially manpower

Standardized Learning

- An international education system, especially difficult for developing countries

- Students differ in gender, ethnicity, academic background, interests, etc.

- => improve the education quality at all levels, emphasize individual talents while maintaining a rigorous standard.

Infrastructure Development

- A major problem in developing countries. Many schools lack adequate school infrastructure.
- In rural areas, the conditions of education are still very poor.
- Implementation of government policies are not enforced in these rural areas, resulting in an aggregation of the problem.
- The quality of many textbooks in the market are not satisfactory and the costs are very high.

Evaluation and Assessment

- Individual (Student) evaluations conducted by instructors as well as institutional evaluation conducted by higher authorities.

- Inflexible methodology and uncertainty make it difficult to develop a standardized system for evaluation.

- As more and more resources are available online, plagiarism problem is more serious than ever before.

Plagiarism Detection

turnitin

Plagiarism Detection in Natural Language

- Paper: Paul Clough, 2003. Old and new challenges in automatic plagiarism detection. National Plagiarism Advisory Service.

- Typical descriminators signaling plagiarism: use of advanced/technical vocabulary, a large improvement in writing style, inconsistencies within the text, a large degree of similarity between two or more texts, shared spelling mistakes or errors

- Multiple texts and single text

Workflow for Plagiarism Detection

discriminators of plagiarism which can be quantified

methods to compare those discriminators

measures of similarity

Plagiarism Detection using N-gram

- Measuring similarity between multiple texts

- Finding the overlap of matching subsequences and substrings (consecutive tokens) of length >= n

- Assumption: the longer n becomes, the more unlikely it is that the same sequence of n tokens will appear in the same order in independently-written texts

N (words)	N-gram occurrences (tokens)	Distinct n-grams (types)	% distinct n-grams	% distinct n-grams in 1 file
1	137204	14407	11	39
2	248819	99682	40	67
3	248819	180674	73	82
4	257312	214119	85	90
5	251429	226369	90	93
6	250956	231800	92	94
7	250306	234600	94	95
8	249584	236310	95	96
9	248841	237409	95	97
10	289610	278903	96	97

Table 1 Uniqueness of consecutive n-word sequences (n-grams) as *n* increases from 1-10 words

$$\text{Similarity}(S_A, S_B) = \frac{\sum_{i \in T} \text{len}_i \times \log(\text{len}_i + 1)}{|S_A|}$$

- Summing elements of the intersection, T, of n-gram sets using a weighted function dependent on the length of the n-gram i.

- Problem: data sparseness, longer n-grams are increasingly rare in derived texts as relatively simple lexical and syntactic rewriting is used.

Improvement: Greedy String Tiling

- An algorithm computing a 1:1 mapping between the tokens in a text pair in such a way that as much of one text is covered with maximal non-overlapping substrings (tiles) from the other.

- Different measurements: minimum and maximum tile length, the average tile length, the dispersion of tile length, and a similarity score based on tile length

Further Improvement (relax the matching between sequences)

- small gaps to represent token deletion
- simple word substitution (using WordNet)
- the insertion of certain words such as domain-specific terminology and function words
- simple re-ordering of tokens

AUTOMATED ESSAY SCORING

With NLP and Machine Learning Applications

Automatic Essay Grading

- Paper: Leah S. Larkey, 1998. Automatic essay grading using text categorization techniques. In the Proceedings of the 21st annual international ACM SIGIR Conference on Research and Development in Information Retrieval.

- Initial Approach: essay length, average word length, multiple linear regression

- Algorithms: Bayesian independence classifier, k-nearest-neighbor classifier, linear regression
- Experimental Data: social studies, physics, a legal argument

	Train	Test	Grades
Soc	233	50	4
Phys	586	80	4
Law	223	50	7

Bayesian Independence Classifiers

- Distinguish better essays from worse essays, dividing the set at different points.

- The log probability that the essay Doc belongs to the class of "good" documents, C, logP(C | Doc) is:

$$\log(P(C)) + \sum_i \begin{cases} \log \left(P(A_i|C)/P(A_i)\right) \\ \quad \text{if the test doc has feature } A_i \\ \log \left(P(\overline{A}_i|C)/P(\overline{A}_i)\right) \\ \quad \text{if the test doc does not have } A_i \end{cases}$$

Feature Selection

- First, remove 418 stop-words from the feature set.
- Remaining terms are stemmed.
- Any stemmed terms found in at least three essays in the positive training set are feature candidates.
- For each binary classifier, calculate Expected mutual information for each feature and rank the features according to the score. The number of features chosen is a tuning parameter for the classifier.

K-nearest-neigbor Classifier

- Find the k essays in the training collection that are most similar to the test essay. The test essay then receives a score which is a similarity-weighted average of the grades of these k essays.

- Using Inquery retrieval system, a probabilistic retrieval system to calculate similarity.

- k is the hyperparameter for tuning.

Linear Regression Features

- 11 features, including the number of characters in the document, the number of words, the number of sentences, average word length, average sentence length, etc.
- Add the results produced by Bayesian classifier and K-nearest-neighbor classifier as additional two features
- Use the SPSS stepwise linear regression package

Variable	Exact	Adjacent	r	Components
Text	.56	.94	.73	BW6, Rootwds Wordlen
Knn (45)	.54	.96	.69	
B1 (200)	.58	.94	.71	
B2 (180)	.66	1.00	.77	
B3 (140)	.60	1.00	.77	
B4 (240)	.62	.98	.78	
All Bayes	.62	1.00	.78	B2, B3
All	.60	1.00	.77	Sents, B2,B3

Table 2: Results on *Soc* data set

Variable	Exact	Adjacent	r	Components
Text	.47	.91	.56	Sents, Wordlen Rootwds
Knn (55)	.44	.90	.53	
B1 (320)	.51	.90	.61	
B2 (480)	.50	.89	.59	
B3 (420)	.55	.90	.63	
B4 (240)	.49	.89	.61	B1,B3
All Bayes	.50	.89	.63	B1,B3
All	.47	.93	.59	B2,B3,B4 BW7,Diffwds, Wordlen Rootwds

Table 3: Results on *Phys* data set

Variable	Exact	Adjacent	r	Components
Text	.24	.66	.57	Rootwds
Knn(90)	.40	.66	.61	
B1 (50)	.36	.54	.60	
B2 (120)	.32	.72	.75	
B3 (300)	.28	.72	.74	
B4 (300)	.28	.84	.76	
B5 (120)	.36	.82	.76	
B6 (160)	.42	.86	.79	
B7 (160)	.32	.78	.78	
All Bayes	.32	.84	.79	B2,B3,B6
All	.36	.84	.77	B2,B3,B6, Knn,BW6

Table 4: Results on *Law* data set

Author's Conclusion

- high value for k in KNN algorithms, and the results showed the KNN approach to be distinctly inferior to both the other approaches.

- Binary classifiers which attempted to separate "good" from "bad" essays produced a successful automated essay grader.


5256 7 1
+ 44697
54-52=
356×356=
2²+4²=
128×128
-55
45
x+y
gettyimages
Imgorthand

Mathematical Word Problem Generation using NLP

- Paper1: Oleksandr Polozov, Elearnor O'Routke, Adam M. Smith, Luke Zettlemoyer, Sumit Gulwani, Zoran Popovic, 2015. Personalized Mathematical Word Problem Generation. In the Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence.

- Paper2: Paul Deane, Kathleen Sheehan, 2003. Automatic Item Generation via Frame Semantics: Natural Language Generation of Math Word Problems. In the Proceedings of the Annual Meeting of the National Council of Measurement in Education.

Requirements for the Generator

- Automatic: a mathematical model, a plot, and a discourse of a word problem are generated automatically from general specifications
- Personalized: students can set preferences for a problem
- Sensible: coherence in a synthesized plot using discourse tropes
- Fit for scaffolding: varying requirements to different layers of a word problem

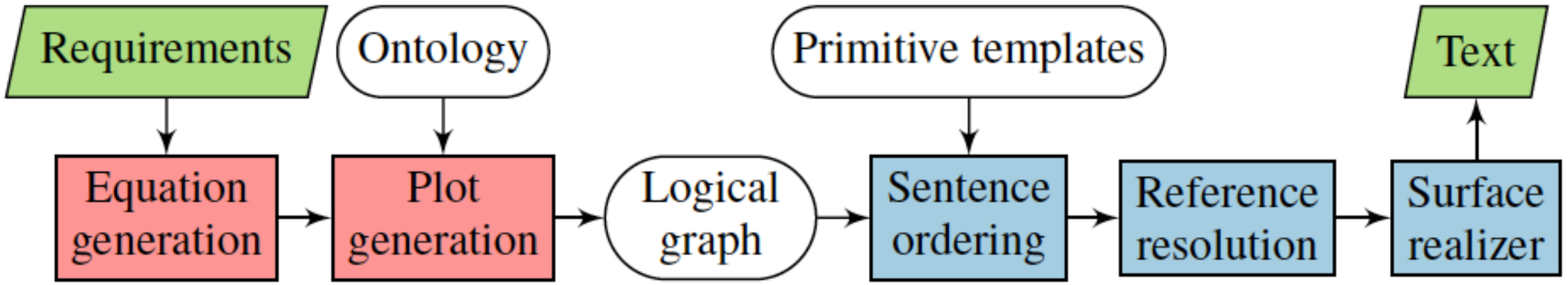


Figure 1: Word problem generation. Red blocks represent logic generation steps, and blue blocks represent NLG steps.

Architecture

- Logic Generation: given the requirements R, it builds the mathematical and the narrative layers of a word problem. The result is a logical graph of actors, actions and entities.

- Natural Language Generation: takes a generated logical graph, and realizes it into a concrete textual representation.

Q1: How comprehensible is the problem? How well did you understand the plot?
Q2: How logical/natural is the sentence order?
Q3: When the problem refers to an actor (e.g. with a pronoun, a name), is it clear who is being mentioned?
Q4: Do the numbers in the problem fit its story (e.g. it wouldn't make sense for a knight to be 5 years old)?

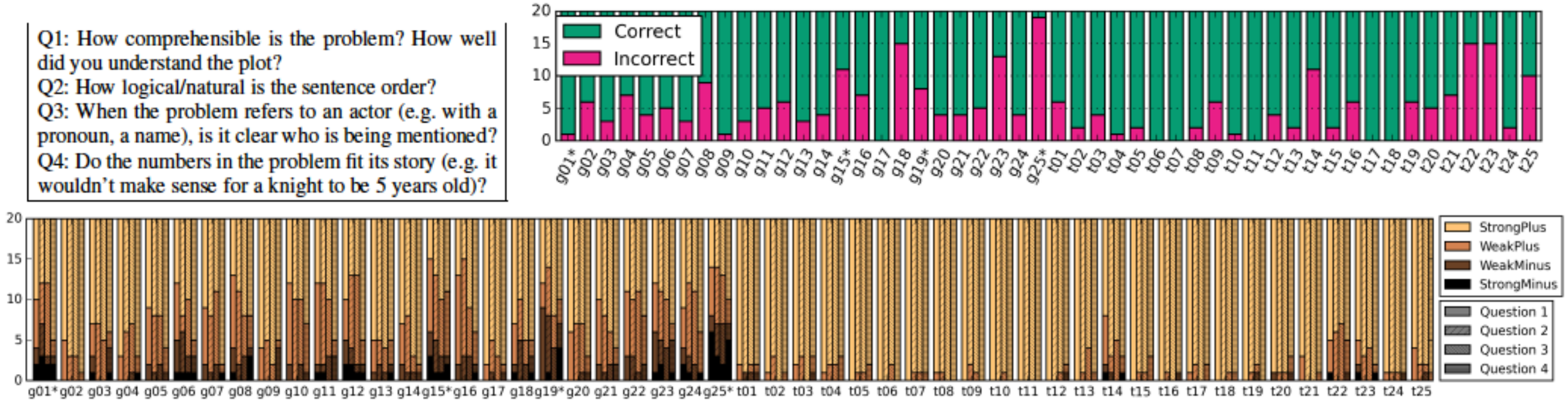


Figure 3: *Top-left:* the forced-choice Likert scale questionnaire for the study A. *Bottom and top-right:* evaluation results for the studies A and B, respectively. Generated/textbook problems are prefixed with g/t. The outliers are marked with an asterisk.

Start: $R_T = \{(? \times 3) - ?\}$, R_S as in Example 2	**Step 1:** $R_T = \{(? \times 3) - ?\}$, R_S: "adversaries" → "friends"	**Step 2:** $R_T = \{(? \times 3) + ?\}$, R_S as in Step 1	**Step 3:** $R_T = \{(? \times 3) + ?\}$, R_S: "Fantasy" → "Wizardry"	**Step 4:** $R_T = \{(? \times x) + ?\}$, R_S as in Step 3
Duchess Alice leads **3 squads** of **12 mounted knights** each in a brave attack upon **duke Elliot**'s camp. Scouts have reported that there are **17 mounted knights** in **his** camp. How many more **mounted knights** does **Elliot** need?	**Duchess Joan**'s countryside consists of **11 towers**, surrounded by **3 villages** each. **She** and **baron Elliot** are at war. **He** has already occupied **16 villages** with the help of **wizard Alice**. How many **villages** are still unoccupied by **Elliot**?	**Orc Bob** has **11 chests**. Inspired by recent advances in burglary, **dwarf Alice** steals **chests** from **the orc**. **They** have **3 gold bars** each. **She** gets a honorable reward of **15 gold bars** from **the master thief Elliot**. How many **gold bars** does **the dwarf** have?	**Professor Alice** assigns **Elliot** to make a **luck potion**. **He** had to spend **9 hours** first reading the recipe in the textbook. **He** spends **several hours** brewing **11 portions** of **it**. **The potion** has to be brewed for **3 hours** per portion. How many **hours** did **Elliot** spend in total?	**Professor Elliot** assigns **Alice** to make a **sleep potion**. **She** had to spend **5 hours** first reading the recipe in the textbook. **It** has to be brewed for **9 hours** per portion. **She** spends **several hours** brewing **several portions** of **it**. The total time spent was **59 hours**. How many **portions** did **Alice** make?

Figure 4: A series of independently generated word problems. For demonstration purposes, each step makes a change of a single aspect in the requirements. The last step demonstrates an unknown variable requirement. Entity references are highlighted.

Summary

- NLP techniques have various applications in the domain of education, including plagiarism detection, automatic essay grading and math problem generation.

- These applications can address, or partially address several challenges in the field.

- Mostly, NLP applications help ease the burden of instructors. For example, automatic essay grading technique allows the instructor to spend more time doing research and developing new course material rather than evaluating students' performance.

- However, higher level problems, including the education system reform and difference between schools cannot be addressed using NLP.

References

- Challenges in Education System. Retrieved January 28, 2018, from http://www.technofunc.com/index.php/domain-knowledge/education-domain/item/challenges-in-education-system
- Education. (2018, January 26). Retrieved January 28, 2018, from https://en.wikipedia.org/wiki/Education
- Leah S. Larkey, 1998. Automatic essay grading using text categorization techniques. In the Proceedings of the 21st annual international ACM SIGIR Conference on Research and Development in Information Retrieval.
- Oleksandr Polozov, Elearnor O'Routke, Adam M. Smith, Luke Zettlemoyer, Sumit Gulwani, Zoran Popovic, 2015. Personalized Mathematical Word Problem Generation. In the Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence.
- Paul Clough, 2003. Old and new challenges in automatic plagiarism detection. National Plagiarism Advisory Service.
- Paul Deane, Kathleen Sheehan, 2003. Automatic Item Generation via Frame Semantics: Natural Language Generation of Math Word Problems. In the Proceedings of the Annual Meeting of the National Council of Measurement in Education.