The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

# Proposition Entailment in Educational Applications Using Deep Neural Networks

**Florin Bulgarov, Rodney Nielsen**
University of North Texas
1155 Union Circle, Denton, Texas, 76203, USA
FlorinBulgarov@my.unt.edu
Rodney.Nielsen@unt.edu

### Abstract

To have a more meaningful impact, educational applications need to significantly improve the way feedback is offered to teachers and students. We propose two methods for determining propositional-level entailment relations between a reference answer and a student's response. Both methods, one using hand-crafted features and an SVM and the other using word embeddings and deep neural networks, achieve significant improvements over a state-of-the-art system and two alternative approaches.

## Introduction

Recent advancements in machine learning have started to put their mark on educational technology. Although the vast majority of the classrooms around the world look essentially the same as they have for several decades, many teachers and students have started to embrace the advantages that technology can bring to the learning process. Moreover, extensive studies have already shown that students who use Intelligent Tutoring Systems outperform students from regular classes (Kulik and Fletcher 2016). This paper focuses on increasing the learning gains in classrooms through algorithms that enhance the analysis between the teacher's reference answer, a student's response, and the relations between them. We contribute by proposing two fine-grained approaches that predict entailment relations between a student's response and each proposition or clause from the teacher's answer. Both methods, one that uses neural networks with word embeddings and the other an SVM model with hand-crafted features, reach similar average $F_1$-scores, significantly outperforming a state-of-the-art system and two alternative approaches. On this account, we make use of Minimal Meaningful Propositions (MMPs) (Godea, Bulgarov, and Nielsen 2016). MMPs have recently been introduced as a decomposition of text into the set of propositions that individually represent single minimal claims or arguments that cannot be further decomposed without losing contextual meaning. By splitting the instructor's reference answer into MMPs, we can make more meaningful comparisons between the learner's response and the individual claims expressed in the reference answer.

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

## Data

We use a modified version of the dataset introduced by Godea, Bulgarov and Nielsen (2016), which contains real-classroom questions, each associated with a teacher input reference answer and an average of 22 student responses. The most important difference is adding the entailment labels. Two graduate students from the Education and Linguistics Department established the proper entailment relations between each pair of reference answer MMP and student response – *understood* (31%) or *not understood* (69%), with a third annotator acting as an adjudicator. A total of 20,815 entailment instances resulted, which were split into train (60%), development (20%) and test (20%) sets.

## Classification

A first approach to this task uses hand-crafted features. The features are split into *general features* and *facet* (Nielsen, Ward, and Martin 2009) *features*. The 45 general features describe general relations between the reference answer MMP and the student response, such as the overall similarities and dependencies between words, Pointwise Mutual Information (PMI) scores, overlapping content, BLEU score, etc. For the rest of the features, we make use of facets due to their granularity level, allowing us to pinpoint the main relations between the two texts. Governor, modifier and relation features are used for each of the following facets: (1) the least likely understood; (2) the most likely understood; and (3), as the averages of all facets. The least and most likely understood facets are chosen by averaging the PMI value between the governors and modifiers of a reference answer MMP facet and all student response's facets. Our second approach to this task is using GloVe word embeddings with a deep neural network (DNN). Specifically, we computed the average embedding vector for each text (reference answer MMP and student response), and combined them into a single vector by concatenating the element-by-element product vector and absolute difference vector (thus, experiments with 200-dimensional word embeddings resulted in a 400-dimensional input to the neural network). The DNN has two hidden layers, each having 64 hidden nodes and a dropout rate of 0.5. Only a small number of iterations was needed to reach the reported results (10 to 20).

| *Model* | $F_1$-*score* *Understst.* | *Not Understst.* | *W. Avg.* |
|---|---|---|---|
| Majority Baseline | 0 | 0.82 | 0.58 |
| LSA | 0.44 | 0.75 | 0.66 |
| Corley and Mihalcea | 0.43 | 0.76 | 0.66 |
| Horbach et al. | 0.39 | **0.83** | 0.67 |
| SVM – man. ftrs. | 0.53 | **0.83** | 0.73 |
| WEs – 50 dim. | **0.63** | **0.83** | **0.76** |
| WEs – 100 dim. | 0.60 | 0.80 | 0.73 |
| WEs – 200 dim. | **0.63** | 0.80 | 0.74 |

Table 1: MMP Entailment Results (WEs = word embeddings, LSA = Latent Semantic Analysis)

## Results

Table 1 shows results obtained by a majority baseline, Latent Semantic Analysis (LSA), and Corley and Mihalcea's (2005) unsupervised system for measuring the semantic similarity of texts. For the latter two approaches, a score was obtained for each pairing of a reference answer MMP and a student response, that was compared against a threshold $t$ estimated on the development set (LSA: $t$ = 0.5; Corley and Mihalcea: $t$ = 0.6). A state-of-the-art system, proposed by Horbach et al. (2013), was also tested for a more meaningful comparison. Even though their system was slightly altered to be applicable to our dataset, the main features remained unchanged.

As can be seen, the approach using word embeddings with 50 dimensions achieves the highest weighted average $F_1$-score of 0.76, performing about 13% better than the state-of-the-art system and the two alternative approaches. The SVM model using hand crafted features obtained a close $F_1$-score, of 0.73. However, we can observe important differences on the *understood* class where word embeddings models achieve a significantly higher $F_1$-score of 0.63. This is a notable improvement of about 43% over just using LSA, which only obtained an $F_1$-score of 0.44. In comparison with the state-of-the-art system, our approach is seeing an increase of 61% on the *understood* class. On the *not understood* class, the difference in results between the alternative approaches and our proposed methods is significantly lower, or none in the case of Horbach et al.'s approach. This is mainly due to the effectiveness of classifying instances in this class utilizing only the word overlap, which is generally very low for the *not understood* class.

**Further experimentation**. Since the manual features and word embeddings (WEs) are fairly independent of each other, combining them should, in theory, further improve the results. As can be seen by comparing rows 1 and 5 in Table 2, SVM achieves a substantially higher $F_1$-score on the *understood* class using WEs instead of the manual features. In fact, adding WEs to our best SVM approach (row 2) results in a weighted average $F_1$-score of 0.76, which is equal to that of the deep neural network (DNN), in row 6. In contrast, inputting the hand-crafted features to the DNN (row 4), substantially decreases the results on the *understood* class. Experiments were also performed where the weights for WEs and manual features were separately learned by individual DNNs and merged at a later stage into a third DNN. These results did not exceed those obtained in rows 2 and 6. A conclusion that can be drawn from these experiments, consistent with what we initially saw in Table 1, is that WEs are more helpful than manual features, particularly in identifying the minority *understood* class. Moreover, even though they seem independent of each other, combining manual features and WEs does not offer much improvement.

| *No.* | *Model* | $F_1$-score *Understst.* | *Not Understst.* | *W. Avg.* |
|---|---|---|---|---|
| 1 | SVM (WEs) | 0.6 | 0.83 | 0.75 |
| 2 | SVM (WEs + man. ftrs.) | 0.61 | 0.83 | 0.76 |
| 3 | DNN (man. ftrs.) | 0.50 | 0.82 | 0.72 |
| 4 | DNN (WEs + man. ftrs.) | 0.55 | 0.82 | 0.73 |
| 5 | **SVM (man. ftrs.)** | 0.53 | 0.83 | 0.73 |
| 6 | **DNN (WEs)** | 0.63 | 0.83 | 0.76 |

Table 2: MMP Entailment Experimental Results (WEs - Word Embeddings 50 dim, DNN = Deep Neural Networks)

## Conclusion

This paper makes use of Minimal Meaningful Propositions in order to break down complex structures and to perform a fine-grained analysis of student responses. This is the first work to do so in a fully automatic process. The two presented approaches exceed the performance of a state-of-the-art and two alternatives systems by approximately 15%, achieving a weighted average $F_1$-score of 0.76.

## Acknowledgements

Research supported by the Institute of Education Sciences, U.S. Department of Education, Grant R305A120808 to UNT. Opinions expressed are those of the authors.

## References

Corley, C., and Mihalcea, R. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment*, 13–18. Association for Computational Linguistics.

Godea, A.; Bulgarov, F.; and Nielsen, R. 2016. Automatic generation and classification of minimal meaningful propositions in educational systems. In *Coling 2016*.

Horbach, A.; Palmer, A.; and Pinkal, M. 2013. Using the text to evaluate short answers for reading comprehension exercises. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 1, 286–295.

Kulik, J. A., and Fletcher, J. 2016. Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of Educational Research* 86(1):42–78.

Nielsen, R. d.; Ward, W.; and Martin, J. h. 2009. Recognizing entailment in intelligent tutoring systems*. *Nat. Lang. Eng.* 15(4):479–501.