Assessing Student Paraphrases Using Lexical Semantics and Word Weighting

Vasile RUS [a,1], Mihai LINTEAN [a], Art GRAESSER [b]. and Danielle MCNAMARA [b]
[a] *Department of Computer Science, The University of Memphis, USA*
[b] *Department of Psychology, The University of Memphis, USA*

Abstract. We present in this paper an approach to assessing student paraphrases in the intelligent tutoring system iSTART. The approach is based on measuring the semantic similarity between a student paraphrase and a reference text, called the textbase. The semantic similarity is estimated using knowledge-based word relatedness measures. The relatedness measures rely on knowledge encoded in WordNet, a lexical database of English. We also experiment with weighting words based on their importance. The word importance information was derived from an analysis of word distributions in 2,225,726 documents from Wikipedia. Performance is reported for 12 different models which resulted from combining 3 different relatedness measures, 2 word sense disambiguation methods, and 2 word-weighting schemes. Furthermore, comparisons are made to other approaches such as Latent Semantic Analysis and the Entailer.

Keywords. intelligent tutoring systems, natural language processing

Introduction

This paper addresses the challenging task of assessing student input in natural language intelligent tutoring systems. In particular, we focus on evaluating student input in iSTART (Interactive Strategy Training for Active Reading and Thinking; [8,9]), an ITS that provides students with reading strategy training. One of the modules in iSTART focuses on training students to paraphrase sentences in science text, called the *textbase* (T). Assessing the student paraphrases (SP) is a critical step in iSTART because it is based on this assessment that the tutoring system could detect possible student misunderstandings and provide the necessary corrective feedback.

An example of a textbase and student paraphrase (reproduced as typed by the student) in iSTART is provided below (from the User Language Paraphrase Challenge [7]):

T: *During vigorous exercise, the heat generated by working muscles can increase total heat production in the body markedly.*

SP: *alot of excercise can make your body warmer.*

The challenge is to automatically decide whether the SP is a paraphrase of the textbase T.

[1]Corresponding Author: Vasile Rus, The University of Memphis, Memphis, TN, 38120, USA; E-mail: vrus@memphis.edu .

In this paper, the student input assessment problem is mapped onto a text-to-text relatedness problem. We describe two methods for detecting text-to-text relations between texts such as paraphrases. In one method, we compute a semantic concept overlap score by greedily matching each concept in the textbase with the most related concept, according to a word-to-word relatedness measure, in the SP. In a second method, concepts in the textbase are weighted by their importance which is estimated using their specificity. While we present the methods in the context of assessing SPs in iSTART, they are generally applicable to other texts. In addition, we compare the proposed methods to other approaches, namely Latent Semantic Analysis (LSA;[6]) and the Entailer[13,15]. LSA, a statistical approach, represents texts based on latent concepts which are automatically derived from an analysis of large collection of texts. The basic idea in LSA is that words that co-occur frequently in similar contexts are semantically related. The Entailer is an approach that relies on both lexical and syntactic information to detect text-to-text semantic relations among sentences. The Entailer, a symbolic approach, proved to be quite successful to assess text relatedness [13,14].

1. Background

Interactive Strategy Training for Active Reading and Thinking (iSTART) is a web-based application that provides young adolescent to college-aged students with self-explanation and reading strategy training [8]. Although untutored self-explanation - that is, explaining the meaning of text to oneself - has been shown to improve text comprehension [2], many readers explain text poorly and gain little from the process. iSTART is designed to improve students' ability to self-explain by teaching them to use reading strategies such as comprehension monitoring, making bridging inferences, and elaboration. Here, we focus on responses in one of the iSTART modules in which the student is asked to only paraphrase the text. Hence, our task is to distinguish good from poor paraphrases. To do so requires capturing some sense of both the meaning and quality of the student paraphrases. Latent Semantic Analysis (LSA; [6]) has been used and studied as an important component in that process. In this paper, we provide an alternative solution based on word relatedness measures computed from knowledge-based resources such as WordNet. We are interested in finding out how well the text-relatedness measures can help detect text-to-text semantic relations as compared to LSA and the Entailer.

1.1. Previous Work

There has been a renaissance recently with respect to exploring computational approaches to detecting text-to-text semantic relations. The recent developments were driven primarily by the creation of standardized data sets for the major relations of entailment (RTE; Recognizing Textual Entailment corpus, [4]), paraphrasing (MSR; Microsoft Research Paraphrase corpus, [3]), and more recently for elaboration (ULPC, User Language Paraphrase Challenge, [7]). Text B is said to be an *elaboration* of text A if text B elaborates on the main topic of text A, i.e. adds something new. Text B is entailed by text A if A logically infers B. We say A and B are in an *entailment* relation. Two texts A and B are in a *paraphrase* relation if and only if they express the same meaning. We focus next on the ULPC corpus as we use it in our experiments.

The ULPC data set comprises annotations for all three types of relations (elaboration, entailment, and paraphrase) as compared to MSR and RTE which only focus on one relation. The corpus contains 1998 textbase-SP pairs collected from previous student-iSTART sessions. The pairs are evaluated on 10 dimensions including entailment and paraphrase quality but also on other quality dimensions such as garbage, i.e. incomprehensible input. These other quality dimensions are not text-to-text relations but rather characteristics of a single text, i.e. the SP. The texts in the MSR and RTE data sets, collected from news articles written by professionals, are grammatically correct with almost no spelling errors and, importantly, with many named entities, e.g. *Mexico*. On the other hand, ULPC texts represent high school students' attempts to self-explain textbases. The student paraphrases are less grammatical, with a relatively large number of misspellings, and no named entities. These characteristics of the ULPCs corpus make it special in some sense and explain some of the choices we made. For instance, we do not use a named entity recognizer in our methods.

2. Approach

Our approach to detecting text-to-text relations relies on word relatedness measures. The word relatedness measures use lexico-semantic information in WordNet to decide semantic similarity between words. WordNet groups words that have the same meaning, i.e. synonyms, into *synsets* (synonymous sets). For instance, the synset of *{affectionate, fond, lovesome, tender, warm}* corresponds to the concept of *(having or displaying warmth or affection)*, which is the definition of the concept in WordNet. Each concept has attached to it a gloss, which contains its definition and several usage examples. Words can belong to more than one synset/concept in WordNet, in case they have more than one meaning. Concepts are linked via lexico-semantic relations such as *hypernymy (is-a)*, *hyponymy (reverse is-a)*, and *meronymy (part-of)*. The nouns and verbs are organized into a hierarchy using the hypernymy relation. A snapshot of the WordNet hierarchy is shown in Figure 1. WordNet contains only content words: nouns, verbs, adjectives, and adverbs.

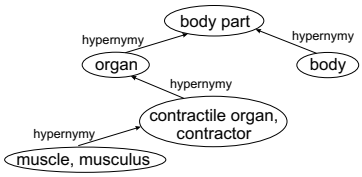

Figure 1. A snapshot of the WordNet taxonomy for nouns.

In general, two concepts are semantically more related if they are closer to each other in the WordNet web of concepts. In Figure 1, the concept of *{muscle, musculus}* is more related to the concept of *{contractile organ, contractor}* than to *{body}*. The rationale is that the former two concepts are one *hypernymy* link away whereas the latter two are four links away (including one change of direction while following the *hypernymy* link beteween *{ body part}* and *{body}*).

There are nearly a dozen WordNet-based similarity measures available [12]. These measures are usually divided into two groups: similarity measures and relatedness mea-

sures. The similarity measures are limited to within-category concepts and usually they work only for the nouns and verbs categories. The text relatedness measures on the other hand can be used to compute similarity among words belonging to different categories, e.g. between a noun and an adjective. The cross-category applicability is very important to us because, for instance, the semantic similarity between the adjective *warmer* in the SP and the noun *heat* in the textbase of the example given in the *Introduction* section can be computed with relatedness measures but not with similarity measures. Therefore, we focus in this paper on the relatedness measures.

2.1. Word Relatedness Measures

We use the following word relatedness measures (implemented in the WordNet::Similarity package [12]): HSO [5], LESK [1], and VECTOR [11]. Given two WordNet concepts, these measures provide a real value indicating how semantically related the two concepts are. We denote with $wn - rel(v, w)$ a generic relatedness function between concepts v and w which would mean any of the three relatedness measures.

The HSO measure is path based, i.e. uses the relations between concepts, and assigns direction to relations in WordNet. For example, the *is-a* relation is upwards, while the *has-part* relation is horizontal. The LESK and VECTOR measures are gloss-based. That is, they use the text of the gloss as the source of meaning for the underlying concept.

One challenge with the above word-to-word relatedness measures is that they cannot be directly applied to compute similarity of larger texts such as sentences. We describe below two methods to extend the word-to-word (W2W) relatedness measures to text-to-text (T2T) relatedness measures. The basic idea of the two methods is to compute the relatedness between the textbase T and SP by averaging over how close individual concepts in the textbase T are to the student-articulated concepts in the SP. Another challenge is the fact that texts express meaning using words and not concepts. To be able to use the word-to-word related measures we must map words in sentences to concepts in WordNet, i.e. we are facing with a word sense disambiguation (WSD) problem. It is beyond the scope of our investigation to fully solve the WSD problem, one of the hardest in the area of Natural Language Processing. Instead, we address the issue in two ways: (1) map words in the textbase T and SP onto the concept corresponding to their most frequent sense, which is sense #1 in WordNet, and (2) map words onto all the concepts corresponding to all the senses and take the maximum of the relatedness scores for each pair of senses. We label the former method as ONE (sense one), whereas the latter is labeled as ALL (all senses).

2.2. Methods

Method 1. In this method, a T-SP relatedness score, $score(T, SP)$, is computed by taking the average of the best W2W relatedness scores between a textbase word and any word in the SP (see Equation 1). The best score between a T-word and a SP-word is found by first computing the similarity between the T-word and all the SP words and then taking the maximum. For words that have a direct match in the SP we assign the maximum relatedness score, which is 1 after normalization.

Method 2. This second method differs from the previous one in that each word in the textbase is weighted by its importance, which is approximated by its specificity. That

$$score(T, SP) = \frac{\sum_{v \in T} max_{w \in SP}\{wn - rel(v, w)\}}{|T|} \quad (1)$$

is, more specific terms in the textbase are weighted more. The specificity of a word is estimated using inverse document frequency (idf). The fewer documents a word occurs in from a large collection of documents the more specific the word is. Idf is computed by inverting the document frequency (df) of the word. Df is derived from Wikipedia, the online encyclopedia. Because of the size (2,225,726 English documents at the time when we processed it) and diversity of topics in Wikipedia, it is assumed that the derived idf values are very good estimates of the true idfs of words.

$$score_{weighted}(T, SP) = \frac{\sum_{v \in T} idf(v) * max_{w \in SP}\{wn - rel(v, w)\}}{\sum_{v \in T} idf(v)} \quad (2)$$

3. Experimental Setup and Results

We used in our experiments the 1998 pairs of Textbase-SP in the User Language Paraphrase Corpus [7]. The performance of the proposed methods is reported along six of the ten dimensions of analysis available in the ULPC: elaboration, semantic completeness, entailment, lexical similarity, and paraphrase quality. It should be noted that some of these dimensions have meanings in ULPC that need be specified as they are not obvious or differ from definitions used by others. In ULPC, elaboration refers to SPs regarding the theme of the textbase rather than a restatement of it. Semantic completeness refers to a SP having the same meaning as the textbase, regardless of word- or structural-overlap. This dimension is of most interest to us because it best matches our goal of detecting semantic similarities among texts. Paraphrase quality takes into account semantic-overlap, syntactical variation, and writing quality. Given these definitions, the semantic completeness dimension in ULPC is equivalent to the paraphrase evaluation in the MSR corpus [3].

In our evaluation, we have explored a space of 3x2x2=12 solutions as a result of combining three relatedness measures (3 - HSO, LESK, and VECTOR), two word sense disambiguation methods (ONE and ALL), and the two methods (with and without IDF-weighting) proposed for extending the word relatedness measures to work for larger texts. Performance is first reported in terms of correlations between the 12 solutions and human judgments. Human judgments are available in ULPC as 6-point interval rating scheme (1-minimum, 6-maximum) for all dimensions. The correlation values are shown in Table 1 where columns represent the six evaluation dimensions we considered and rows represent different solutions. For instance, the row ALLIDFLESK means a solution that uses ALL the senses of words to compute word-to-word relatedness, weights words using IDF values, and applies the LESK relatedness measure. If no IDF is mentioned in the name of a solution but rather a dash (-), e.g. ONE-LESK, it means no word weighting was used. We also show correlations for LSA, and three variants of the Entailer, which

are reported in the ULPC package. We picked these four other approaches because they best correlated with human judgments on the six dimensions [7].

We also evaluated the 12 solutions in terms of accuracy, which is the percentage of correct predictions out of all predictions. In order to measure accuracy, we used the binary values for human judgments in ULPC (1-3.49 = 0 [low]; 3.50-6 = 1 [high]). The accuracy results were obtained using 10-fold cross-validation. In k-fold cross validation the available data is divided into k equal folds. Then, k trials are run, one for each fold. In each trial one fold is set aside for testing and the other $(k-1)$ are used for training. The average of the accuracies for the k trials is reported. When $k = 10$, we have 10-fold cross-validation. The training consists of finding a threshold value for a particular solution, e.g. ONE-LESK, above which a prediction is considered high, and low otherwise. These predictions are then compared with the binary human judgments in order to compute the accuracy.

An analysis of the results in Table 1 indicates that a combination of IDF weighting with the VECTOR relatedness measure (both for ONE-sense and ALL-senses methods) provides best correlations with humans. For instance, the correlation for ONEIDFVECTOR and ALLIDFVECTOR on the semantic completeness are the highest among all the methods, .603 and .606, respectively. These two solutions are also the best performers for entailment and writing quality. Another interesting finding is revealed by comparing the correlations for models that use IDF weighting and those that do not. Indeed, using IDF weighting helps. For example, the correlations for ONEVECTOR and ONEIDFVECTOR are .567 and .603, respectively, along the semantic completeness dimension. On the other hand, no real benefit is evident from using all the senses of words (ALL) as opposed to using only one sense (ONE) to compute relatedness. Some small improvements are noticeable though. For instance, correlations along the semantic completeness dimension for two models that differ only in their word sense disambiguation method, ONEVECTOR versus ALLVECTOR, show an improvement from .567 to .575.

Accuracy results, not shown or discussed for all the 12 solutions and 6 dimensions due to space reasons, vary within a relatively small range. For instance, along the semantic completeness dimension, the accuracy scores for the 12 solutions vary from 78.12% (for ALL-LESK) to 79.47% (ONE-HSO, ONEIDFLESK, or ALLIDFLESK). LSA provides an accuracy of 77.62% while R-Ent provides 78.47%. A baseline approach that guesses all the time the dominant label in the data provides 69.36% accuracy. We also noticed that along the elaboration dimension the accuracy values are extremely high for all 12 explored solutions, LSA, and the Entailer. This is explained by the distribution of the 1998 pairs along this dimension: 98.29% of the instances are low. Therefore, the baseline method that always guesses low elaboration yields an accuracy of 98.29%. The distribution of the low-high instances along many dimensions is highly biased towards one of the low or high values. The only balanced dimension seems to be the paraphrase quality dimension with 53.90 instances being labeled as high. Due to the biases in the data, kappa coefficient, a measure of agreement between predictions and human judgments that also accounts for chance agreement, vary more. Kappa takes values between -1 and 1, with 1 meaning perfect agreement, 0 no agreement, and -1 perfect disagreement. In our experiments, kappa is nearly 0 (meaning chance agreement) for all the solutions along the elaboration dimension. For the more balanced dimension of paraphrase quality, kappa varies from 33.52 (F-Ent) to 44.93(for IDFONELESK). LSA yields a kappa of 36.14 and the best kappa for an Entailer method is 43.36 (R-Ent).

Table 1. Correlations between different solutions and human judgments. (@ means results are on a subset of 200 pairs; ! - means not significant; * significant at 0.05 level; all other values are significant at 0.01 level)

Method	Elab	Sem-C	Ent	Lex-sim	Par-Q	W-Q
ONE-HSO	-.156	.556	.515	.791	.318	.447
ONE-LESK	-.160	.550	.507	.784	.310	.441
ONE-VECTOR	-.137	.567	.531	.800	.341	.495
ONE-IDFHSO	-.171	.585	.539	.798	.387	.464
ONEIDFLESK	-.179	.576	.526	.792	.374	.456
ONEIDFVECTOR	-.149	.603	.563	.812	.422	.530
ALL-HSO@	-.238*	.459	.422	.752	.081!	.428
ALL-LESK	-.146	.560	.517	.788	.323	.459
ALL-VECTOR	-.110	.575	.541	.791	.362	.529
ALLIDFHSO@	-.233*	.476	.467	.734	.140!	.415
ALLIDFLESK	-.161	.583	.534	.792	.386	.473
ALLIDFVECTOR	-.113	.606	.568	.794	.443	.568
LSA	-.175	.555	.535	.804	.410	.498
R-Ent	-.177	.564	.512	.776	.321	.425
F-Ent	-.212	.449	.441	.726	.269	.405
A-Ent	-.204	.529	.497	.785	.308	.434

3.1. Discussion and Future Work

A qualitative analysis of the various methods we experimented with revealed several important aspects of our evaluation. These aspects may explain some of the errors produced by the proposed methods. *First*, the results we reported were obtained on the raw SP as typed by students, with many typos. These typos lead to failed word relatedness measurements which in turn lead to inaccurate estimates of semantic similarity. It would be interesting to compare the discussed methods on a typos-free set of SP. *Second*, the W2W relatedness measures do not exploit the full potential of WordNet lexico-semantic information as for instance, *good* and *bad* are as similar as *bad* and *evil*. Using VECTOR as the measure, we obtained relatedness values of 0.718 and 0.714 for these pairs of concepts, respectively. This is simply the case because the W2W relatedness measures only account for the number and eventually direction of the links but not the label of the links. Between, *good* and *bad* there is a *antonymy* relation while between *bad* and evil there is a *similar-to* relation. The power of our methods to discover semantic similarity is limited by the power of the WordNet relatedness measures. *Third*, the low kappa values may suggest that a comparison of the various methods on balanced data sets (50-50 split between low and high values) along each of the evaluation dimensions may offer a crisper comparison of the methods. We plan to run such a comparison in the future. It is also worth mentioning that we had difficulties using HSO in combination with ALL senses (see @ in Table 1).

4. Conclusions

We presented in this paper two methods for assessing student self-explanations in the intelligent tutoring system iSTART. The methods rely on word-to-word relatedness mea-

sures and also on word weights derived from Wikipedia. The proposed methods are generally applicable to other types of text although they were presented in the context of iSTART. Word-weighting combined with the VECTOR relatedness measures best correlated with human judgments.

Acknowledgments

This research has been supported in part by funding from the Institute for Education Sciences (IES R305A080589) and National Science Foundation (RI 0836259).

References

[1] Banerjee, S., and Pedersen, T. 2003. Extended gloss overlaps as a measure of semantic relatedness. In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, 805-810.

[2] M.T.H. Chi, N. de Leeuw, M. Chiu, M., and C. LaVancher, Eliciting self-explanations improves understanding, (1994), Cognitive Science, 18, 439-477.

[3] W.B. Dolan, C. Quirk, and C. Brockett, Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In Proceedings of COLING 2004.

[4] I. Dagan, O. Glickman, and B. Magnini, The PASCAL Recognising Textual Entailment Challenge. In Proceedings of the Recognizing Textual Entaiment Challenge Workshop, 2005.

[5] Hirst, G., and St-Onge, D. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Fellbaum, C., ed., WordNet: An electronic lexical database. MIT Press.

[6] T. Landauer, D.S. McNamara, S. Dennis and W. Kintsch (Eds), Latent Semantic analysis: A road to meaning, 2007, Mahwah, NJ:Erlbaum.

[7] P.M. McCarty and D.S. McNamara, User-Language Paraphrase Corpus Challenge, online, 2008.

[8] D.S. McNamara, I. Levinstein, and C. Boonthum, iSTART: Interactive Strategy Trainer for Active Reading and Thinking, Behavioral Research Methods, Instruments, and Computers, 2004, 36 (2), 222-233.

[9] D.S. McNamara, C. Boonthum, I.B. Levinstein, and K. Millis, Evaluating selfexplanations in iSTART: comparing word-based and LSA algorithms. In Landauer, T., D.S. McNamara, S. Dennis, and W. Kintsch (Eds.), Handbook of LSA, Mahwah, NJ: Erlbaum, 2007, 227-241.

[10] G. Miller, Wordnet: a lexical database for english, Communications of the ACM, 1995, 38(11):39-41.

[11] Patwardhan, S. 2003. Incorporating dictionary and corpus information into a context vector measure of semantic relatedness. Master's thesis, Univ. of Minnesota, Duluth.

[12] T. Pedersen, S. Patwardhan, and J. Michelizzi, WordNet::Similarity - Measuring the Relatedness of Concepts, In the Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04), pp. 1024-1025, July 25-29, 2004, San Jose, CA (Intelligent Systems Demonstration)

[13] V. Rus and A. C. Graesser, Lexico-Syntactic Subsumption for Textual Entailment, In Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005, John Benjamins Publishing Company, ISBN: 9789027248077, 2007

[14] V. Rus, M. Lintean, P.M. McCarthy, D.S. McNamara, and A.C. Graesser, 2008, Paraphrase identification with lexico-syntactic graph subsumption, In D. Wilson and G. Sutcliffe (Eds.), Proceedings of the 21st International FLAIRS Conference (pp. 201-206), Menlo Park, CA: The AAAI Press.

[15] V. Rus, P.M. McCarthy, D.S. McNamara, and A. C. Graesser, A Study on Textual Entailment, International Journal on Artificial Intelligence Tools, 17(4):659-685-499, August, 2008