AUTOMATIC ANSWER EVALUATION: NLP APPROACH

Arun P V[1],Parshu Ram Dungyel[2], Karma Wangchuk[3], Kesang Wangmo[4], Uttar K Rai[5], Yeshi Jamtsho[6]

Department of Information Technology, College of Science and Technology, Royal University of Bhutan

[1]arunpv@cst.edu.bt, [2]prsharma@cst.edu.bt,
[3]karmastar17@gmail.com,[4]2010045@cst.edu.bt, [5]2010148@cst.edu.bt,
[6]khengshi@gmail.com

ABSTRACT

Automatic assessment of subjective answers requires Natural Language Processing(NLP)based evaluation and automated assessment. Various techniques used are Ontology, Semantic similarity matching and Statistical methods. An automatic short answer assessment system based on NLP is attempted in this paper. Various experiments performed on a dataset, revealed that the semantic Enhanced NLP(ENLP) method outperformedmethods based on simple lexical matching; resulting upto 85 percent performance with respect totraditional vector-based similarity metric.

Keywords: *Natural Language processing, Keyword analysis, Information Extraction, Semantic matching.*

1. INTRODUCTION

Automatic evaluation is preferred to manual assessment to avoid mono tonic, bias errors and to conserve teacher's time for main activity. Hence automatic assessment is vital for educational system. Computer Assisted Assessment (CAA) is an area of extensive research especially due to larger intake by universities and adoption of e-learning system as ubiquitous education platform as well as due to the developments in Natural Language processing(NLP),Information Extraction (IE) and e-learning. (Metzler, Dumais and Meek (2007); Salton,1989; Willet,1988).

Ontology based methods, Key word analysis, natural-language processing and Information mining techniques(Callan,1994; Willet,1988) are the main approaches adopted for text assessment. Keyword analysis has usually been considered a poor method as it is difficult to tackle problems such as synonymy or polysemy in the student answers, on the other hand, a full text parsing and semantic analysis is hard to accomplish, and very difficult to port across languages. Hence, Information Extraction offers an afford able and more robust approach, making use of NLP tools for searching the texts for the specific contents and without doing an in-depth analysis(Callan,1994).

Methods like combining keyword based methods(Smeaton,1992), pattern matching techniques (Hatzivassiloglou,2001),breaking the answers into concepts and their semantic dependencies (P´erez, Alfonseca and Rodr´ýguez (2004b)), Machine Learning techniques(Papineni,Roukos,WardandZhu(2001)),LatentSemanticAnalysis (LSA) (P´erez, Alfonseca and Rodr´ýguez (LREC-2004)), and LSA with syntactic and semantic information(Banerjee and Lavie (2005); Snow and Vanderwende (2006)) are the other techniques used for the assessment of student's free text answers.

This research envisages the automatic assessment by enhanced NLP method. The enhanced version of NLP based algorithm is explained in section two. The system architecture of ENLP method and metrics for evaluating the quality of an automatics coring algorithm is explained in section three. The Section four illustrates about the experiment at ion performed on the proposed system.

2. THE ENHANCED NLP BASED METHOD

This method assesses a text by computing a score based one xplicit concept match between the student's answer and teacher's answer (i.e. reference). If more than one reference is available, them at ching similarity is scored against each reference in dependently and the best scoring pair is used to find the final score. The concepts are converted to intermediate forms and are matched based on the following modules.

EXACT MODULE: This module matches concepts only if their surface forms match.

STEMMING MODULE: This matches two concepts to each other if they are identical after being passed through the intermediate form generator.

HEURISTICS RULE BASED MODULE: This module maps two concepts to each other if they share the same base form based on some heuristics rules.

RULE 1-WORD NET SYNONYM MATCH:

If intermediate forms of target and references are matched with reference to the synonyms it shows that they both will have same parts of speech and belongs to the same synset in Word Net.

RULE 2-NUMERIC VALUE MATCH:

The numeric value features to each part of text inferred to correspond to a numeric value.(Eg. "7th"is aligned to "seventh").

RULE3- ACRONYMMATCH:

It aligns pairs of node with the properties of capitalize d letters and the letters correspond to the first characters of some multi word.(Eg."NLP" is a ligned with "Natural Language Processing").

RULE4 - DERIVATIONAL FORM MATCH:

This Rule aligns sentences which have the same inherent idea by using predicate rules.

RULE5-COUNTRY ADJECTIVAL FORM /DEMONYM MATCH:

It matches from an explicit list of place names, adjectival forms, and demonyms.(Eg."Chennai" and"Madras")

The steps of Enhanced NLP(P´erez, Alfonseca and Rodr´ýguez (2004b,LREC-2004); Papineni, Roukos, Ward and Zhu(2001)), algorithm are given below.

1. The matching of concept is counted

for each reference.
2. Combines the scores of each reference as the weighted linear averages of marks.
3. The short and irrelevant answers are penalized by a penalty factor.

PREPROCESSING MODULE:

Transforms the student's answers as well as the reference key to intermediate form. The texts are broken in to tokens (e.g. words, numbers and punctuation symbols) and the sentence boundaries as well as the hot spots are identified. The other processes like stemming and stop-words removal are also the part of this module and are summarized in the fig.1.

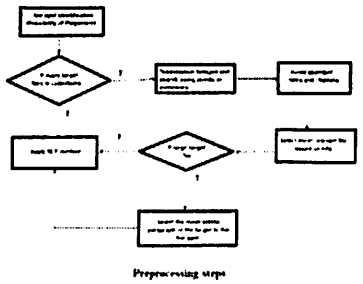

Figure 1: Preprocessing steps

The steps of the enhanced NLP approach are shown in fig.2.

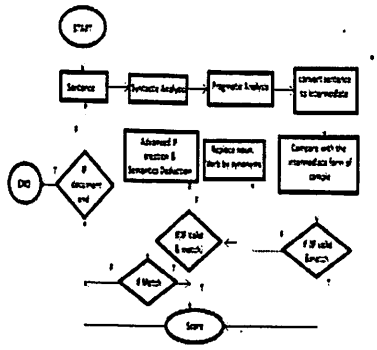

Figure 2.Summarization of NLP method

INTERMEDIATE FORM GENERATION:

The intermediate form production is the basis of this algorithm and this intermediate form should be such that it must be able to uniquely represent the ideas. It is of the form almost similar to the predicate calculus and has the following peculiarities.

1. Voice change detection.
2. Detect tense changes Synonym.
3. Idea similarity detection.

LEARNING PHASE:

This stage deals with yielding the meaning of a sentence which can be utilized to create rules based on which deductions can be made. The sentence meaning is obtained from word meaning stored in lexicon. Internal representation of a sentence is made from the internal representation of words (meaning of words)

E.g.: Meaning of block is: (inst? x block)
Meaning of red is: (color x red)
Then meaning of sentence, Block is red :
(and (inst? x block) (color? x red))

Thus the semantics reveal the contextual knowledge as well as world knowledge to the system which are very relevant for deduction and based on which the rules used for deductions are derived from a given text.

SUMMARIZING PHASE :

The rules and the facts obtained have to be summarized based on their relative meaning as well as transitivity, conditional symmetry. Removal of redundancy etc to give optimization. The facts can be summarized by adopting a binary tree arrangement (for source as well as target files) and also by asking certain automated queries by the system based on the context (Enabling a system

to ask queries means that it had been given some intelligence!).

ADVANCED DEDUCTION:

Deductions are done based on rules mentioned as well as on the contextual knowledge provided. This contextual knowledge analyzed during semantic analysis is used for creating rules and based on which deductions are made. For e.g: If there is a rule such that "two loving person cares each other". Then this can be represented as

l o v e s (x, y) -----> cares(x, y)

(These Rules are obtained during Semantic analysis).

By using which we can deduct that if x loves y then x cares y. We can deduct using the semantic knowledge and can use it to deduct if x loves y then x cares y. We can deduct using the semantic knowledge and use it to deduct certain sentences from the given ones using which we can check whether an idea is variably represented and hence to detect the level 3 matches

VALIDATION MODULE:

In this module, the data sets that we used for the experimentation are evaluated by human judges who are experts in the concerned subjects. In comparison steps, the humans core and systems core are compared and the correlation between human and systems core is computed. The module is summarized in fig.3.

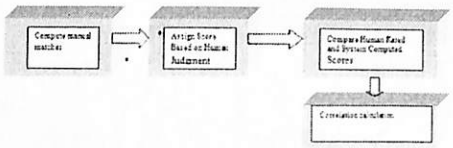

Fig.3 Validation

3. IMPLEMENTATION

The system is implemented in java, prolog with database as DB2 and prolog Db is used for storing the facts (intermediate forms) and rules (obtained from semantic analysis) as shown in fig 4.The syntactic analysis is carried to check the syntactic structure of the sentence and to identify the parts of speech for intermediate form generation. It is also used in semantic analysis phase for the extraction of rules and also for relevance checking in preprocessing module. This phase is implemented using Stanford parser which produces a parse tree which is used to generate intermediate form. Pragmatic analysis replaces the pronouns by their referents and interprets the limited range of metonymy. After a sentence has been parsed if there is a pronoun then it is replaced by the subject or pronoun referent of the previous sentence if previous sentence was simple. If complex sentence exist then priority is considered for resolution. The intermediate form as explained is similar to the predicate calculus for efficient processing due to the inherent deduction of predicate calculus and is generated using output of parser. To check the chance of synonym replacement by the students the intermediate forms are replaced by their synonyms and for this purpose Word net is used. The Word Net data base is accessed through java program and a word in the sample is replaced by its synonym and compared with the intermediate form of the original. The learning phase (semantic analysis) phase is done to derive the rules and is done using word net (to acquire information regarding synonyms, antonyms etc) and Stanford parser (to acquire information about the structure of sentence, nature of a word etc.) and resulting rules are added to prolog db for advanced deduction to check idea match. Summarization phase optimizes the result of learning phase and IF generation phase and is done by asking auto questions, by tree optimization etc. implemented by storing mathematical

relations, sentence structure information, word meaning information etc. The intermediate form is generated for target and sample and both are compared. Deductions on intermediate form for tackling idea are implemented using a prolog db. For this purpose prolog is used and interfacing is done by using JIP and SW-prolog is used for the purpose. The IF of original document is stored as facts in prolog data base and rules of the original are also supplied. Then the IF of the target is supplied as queries to prolog and prolog will do deduction using the rules and hence any match can be easily detected and on a similarity the matching score is incremented. Both source and target files are subjected to the process shown in fig 4.

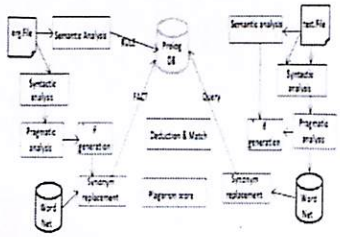

Fig.4. Implementation

4. EXPERIMENTATION & EVALUATION

A bench mark data set released by Rada Mihalcea, Michael Mohler and another created from actual evaluations in our college. The ten sets sum up to a totalof1929student'sanswers and many different alternative keys were provided and evaluated by different each ers that consisted of descriptions, definitions, Yes or No, advantages and disadvantages as themes. The metric used to evaluate the goodness of the free-text scoring of answers to this corpus has been the Pearson correlation filling one vector of scores with the human's scores and the other with the automatic scores. In this way, the algorithm has been evaluated, for each of the data sets, by comparing the concepts from the student answers against the references, and obtaining the final score for each candidate.

TABLE

1.Comparingcorrelationsproducedbydiffe rentmodule stages

Exp.	Mapping Modules	Correlation
1	Level 1(sentence match)	0.80
2	Level 2(synonym deduction)	0.78
. 3	Level 3(semantic deduction)	0.71

5. COMPARISON WITH EXISTING EVALUATION ALGORITHMS

Baseline scoring algorithms used in this work include:

Keywords

consists in looking for coincident keywords or n-grams in any of the reference texts but it cannot deal with synonyms or with polysemous terms in the student answers.

VSM

using avectorial representation of similar answers, we have done a five-fold cross-evaluation, in which 30% of the texts are taken as training set.

ERB

The main principle behind ERB algorithm is the measurement of the overlap in sentences.

Table 2 shows that comparison of ENLP scores with other methods. The first column indicates the scores obtained by ENLP. The other existing methods ERB, keywords and VSM are represented in consecutive columns. The ENLP method gives the better and VSM method

obtained the least score. The ASAGS out performs the other method for 80% of the dataset. It gives the good average correlation compared with other methods.

TABLE

2. Comparison of ASAGS with three other methods

Data Sets	ENLP	ERB	Key words	VSM
1	0.81	0.63	0.12	0.35
2	0.55	0.36	0.23	0.09
3	0.51	0.36	0.19	0.24
4	0.73	0.82	0.57	-----
5	0.51	0.41	0.57	0.52
6	0.35	0.02	-0.05	0.05
7	0.51	0.21	0.32	0.17
8	0.53	0.41	0.22	0.17

The ERB and key word based method gives higher correlation for a few cases.

6. CONCLUSION

Thus we have implemented NLP based answer evaluation system and compared with other methods in literature. Our approach out performs the other methods for 80% of the dataset and the ERB and Keyword based method for a few cases. The parallel implementation of this approach may be further investigated.

REFERENCES

Metzler.D.,Dumais.S..andMeek,C."Similaritymeasuresforshortsegmentsoftext",InProcofECIR, 2007.

Peter Willet. "Recent Trendsin Hierarchical Document Clustering." Information Processing and Management, 24(5):577-597.1988.

Gerard Salton. "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer."Addison-Wesley, Reading, Massachusetts.1989.

Jaime P. Callan. "Passage-Level Evidence in Document Retrieval." In Proceedings of the 17th A CM SIGIR International Conference on Research and Development in Information Retrieval, pages 302-309, Dublin.Ireland.1994.

Alan F. Smeaton. "Progress in the Application of Natural Language Processing to Information Retrieval Tasks." The Computer Journal. 35(3):268-278.1992.

VasileiosHatzivassiloglou, etc. "SimFinder: A Flexible Clustering Tool for Summarization." NAACL Workshop on Auromatic Summarization, Association for Computational Linguistics.2001

P'erez, D, Alfonseca E and I.Rodr'yguez, P. "Upper bounds of the BLEU algorithmapplied to assessing studentessays".In Proceedings of the 30th International Association for Educational Assessment(IAEA)Conference. 2004b.

K.Papineni.S.Roukos,T.Ward12.andW.Zhu. "BLEU: a method for automatic evaluation of machine translation", Technical Report RC22176 (W0109-022), IBM Research Division,ThomasJ.WatsonResearchCenter.2001.

PérezD.AlfonsecaEandRodríguez13.P."Application oftheBleumethodforevaluatingfree-textanswersinane-learningenvironment",InProceedingsofthel.anguage ResourcesandEvaluationConference(LREC-2004). Lisbon,2004.

Banerjee.S.andLavie,A.2005:'METEOR:An Automatic Method for MTE valuation with Improved Correlation with HumanJudgments'.In*ACL2005 Workshop on Intrinsicand Extrinsic Evaluation Measures for MT and/or Summarization*. AnnArbor, MI:pp.65-73.

RionSnow,LucyVanderwende.2006EffectivelyUsingSyntaxforRecognizingFalseEntailment

RadaMihalcea, Michael Mohler, http://www.cse.unt.edu/~rada/downloads.html