Using a MaxEnt Classifier for the Automatic Content Scoring of Free-Text Responses

Jana Z. Sukkarieh

Educational Testing Service, Rosedale Road, Princeton NJ 08541, USA

Abstract. Criticisms against multiple-choice item assessments in the USA have prompted researchers and organizations to move towards constructed-response (free-text) items. Constructed-response (CR) items pose many challenges to the education community - one of which is that they are expensive to score by humans. At the same time, there has been widespread movement towards computer-based assessment and hence, assessment organizations are competing to develop automatic content scoring engines for such items types – which we view as a textual entailment task. This paper describes how MaxEnt Modeling is used to help solve the task. MaxEnt has been used in many natural language tasks but this is the first application of the MaxEnt approach to textual entailment and automatic content scoring.

BACKGROUND, AIM, AND PROBLEM DEFINITION

Criticisms against multiple-choice item[1] assessments in the USA have prompted organizations to move towards constructed-response (free-text) items. Constructed-response (CR) items pose many challenges to the education community - one of which is that they are expensive to score by humans. At the same time, there has been widespread movement towards computer-based assessment and hence, assessment organizations are competing to research and develop automatic content scoring engines for such item types. In theory, the definition of "content" seems to be "what the text says or conveys". In practice, content and text understanding mean different things for different researchers without taking into account the construct of a test item, the particular task at-hand, and the associated validity issues. Content could mean "technical content" in a text as in SEAR software (Christie, 2003) or it could also mean "[the] semantic similarity between pieces of textual information" as in the Intelligent Essay Assessor software (Rehder, 1998). Content might also mean focus, coherence, or organization (Burstein, Marcu, & Knight, 2003). It might mean topical content, vocabulary variance, usage, or vocabulary types in some systems such as BETSY (Rudner and Liang, 2002). To Valenti, Neri and Cucchiarelli (2003) reporting on the Intelligent Essay Marking System (Ming, Mikhailov, and Kuan, 2000), content is restricted to "content-based subjects." This tautological restriction might point to a definition such as 'scientific subject matter' versus 'poetry' but it might also be a symptom of the lack of a definition of content when it comes to automatic content scoring, i.e., "content" seems to mean only what a technology deems it to be. Without

[1] An item is a test question and includes its rubrics, which are a set of instructions about how to score the item.

going into a philosophical debate, in this paper we concentrate on **analytic-based content**. This is the kind of content that is specified by test developers in the rubrics in terms of main points or **concepts**. While scoring each student's response, human raters look for evidence that demonstrates a student's knowledge of the concepts. The following shows an example of a Reading Comprehension test item, with required analytic-based content for the response (as shown by the three concepts C_1, C_2, & C_3) and the recommended rubric for assigning score points.

Reading Comprehension Item (Full credit: 2 points) *Stimulus:* A reading passage *Prompt:* In the space provided, explain why the reading passage suggests that Spring is coming?	*Concepts or Main Points:* **C_1:** Trees only bud in Spring **C_2:** Rain is often seen in the Springtime **C_3:** Robins are only found in the Spring

Scoring rules (0,1,2 are scoring points):

2 Points for C_1
1 Point for C_2 (only if C_1 is not present) or for C_3 (only if C_1 and C_2 are not present)
0 Points otherwise

We view the task we are solving as a **textual entailment (TE)** problem with the additional complexity that students' responses might contain misspellings, ungrammaticality, foreign words, abbreviations, interjections, common short message service-like (SMS) words, or mixed-mode representations of text and equations. We use TE to mean either a paraphrase or an inference (up to the context[2] of the item). The task of analytic-based content scoring is reduced to a TE problem in the following way:

> **Given** a concept, *C*, (e.g., "the body increases its temperature") **and** a student response, *A*, (e.g., either "the body raises its temperature," "the body responded. His temperature was 37° and now it is 38°," or "Max has a fever") **and** the context of the item, **the goal is** to check whether *C* is an inference or paraphrase of *A* (in other words, *A* implies *C* and *A* is true).

The question then is: Does a student's response entail concept C_i? Once this question is answered for each *i*[3], scoring rules are applied and a score is given to the student. In this paper, TE is seen as a classification problem and a maximum entropy modeling (MaxEnt) is used to implement the classifier. MaxEnt has been suggested and used in natural language processing (NLP) by many, for example, (Berger, Pietra & Pietra, 1996), (Ratnaparkhi, 1997), (Mikheev, 1999), (Nigam, Lafferty & McCallum, 1999), (Chieu & Ng, 2002) and (Osborne, 2002). It has been shown to be a competitive text classification algorithm requiring little supervision; as far as we know, this is the first application of the MaxEnt approach to textual entailment and

[2] The context may include the stimulus, reading passages (if any), the rubrics, the subject matter, and the grade level. Any of these factors might affect the scoring.

[3] Note that in this paper, the assumption is that a decision on whether a student response entails a C_i is independent of that decision for C_j where $i \neq j$.

automatic content scoring. The application of MaxEnt is used in **c-rater®,** the Educational Testing Service (ETS) technology for the analytic-based content scoring of short free-text responses written in English. However, the MaxEnt application is neither necessarily restricted to short text nor to English and can be applied in the same way to essay-length responses. In the following, we describe how MaxEnt is used towards the entailment task. Next, we will report results for this technique using items administered in schools in Maine and Massachusetts, USA. We show that even with a set of potentially noisy training data, MaxEnt achieves a higher accuracy than a rule-based approach. Finally, we will conclude with next steps.

MAXENT ENTAILMENT DETECTION

Maximum Entropy

MaxEnt modeling is a probability distribution estimation technique that uses a closed-world principle, i.e., the technique models only that which is known and assumes nothing else. That which is unknown is modeled as if MaxEnt were a uniform distribution. Maximizing the uncertainty, randomness, or entropy ensures the least-biased distribution. This also makes the algorithm resistant to noise (Goldwater & Johnson, 2003). That which is known is usually a set of observations, facts or evidence that one makes about a sample of training data and writes these observations in terms of constraints or feature functions with constraints over their values. The MaxEnt probability distribution has the exponential form (Ratnaparkhi, 1997):

$$p(a\,|\,b)=\frac{1}{Z(b)}\prod_{j=1}^{k}\alpha_j^{f_j(a,b)}$$

Where *a* is the class to predict, *b* is a given context or textual material and *Z (b)* is a normalization function that ensures that $\sum_a p(a\,|\,b)=1$. Each feature function $f_j(a,b)$ is usually binary but this differs from one implementation to another. The parameters α_j are estimated by a procedure called Generalized Iterative Scaling (Darroch & Ratcliff, 1972). We have used the Java-based OpenNLP maximum entropy package (http://maxent.sourceforge.net) where all we need to do for a particular task is to define or provide the set of outcomes or classes a, the set of possible contexts b, a training dataset, Ґ, and the feature functions representing the observations over Ґ.

Entailment Task

In our case, we would like to obtain a probability on whether a student's response entails a concept *C_i*. We will denote Student_Response with *SR* and Response_Sentence with *RS*. We define:

$$prob(Label\,|<S\,\mathrm{R},C_i>)=\max_{\lambda}\{prob(Label\,|<\mathrm{R}\,S_{\lambda},C_i>)\}$$

Where λ is the number of sentences in a student response, C_i is concept *i* and *Label* can be {*1, 0*} for entailment or non-entailment, respectively. MaxEnt algorithm is used to obtain a probability predicting the Label for each argument on the right-hand side of the above equation i.e. does RS_λ entail C_i? The training data, Ґ, from which the MaxEnt algorithm learns its model consists of: (a) approximately 1,300 triples of the form <$Sentence_\alpha$, $Sentence_\beta$, *1*> that have been randomly extracted from students' responses corresponding to around 100 items and annotated manually, and (b) approximately 500 triples of the form <$Sentence_\alpha$, $Sentence_\beta$, *0*> extracted from no-credit student responses corresponding to the same 100 items. No-credit response means that humans judged that no sentence in the response entails any of the concepts in any of the items. Not all lexical entities in $Sentence_\beta$ are weighted equally. The most essential ones are flagged. We refer to these lexical entities as required entities.

It is also possible that a test developer provides more than one alternative for a concept e.g. "Trees bud in the spring" could also be given as "spring is the time for sprouting". If we denote the alternatives of C_i with AC_{ij} where *j* is the number of these alternatives then given the MaxEnt model (learnt using the same training data) and assuming that these alternatives are independent[4] then we define:

$$prob(Label\,|{<}\,S\,\mathrm{R}, C_i >) = \max_{\lambda, j}\{prob(Label\,|{<}\,\mathrm{R}\,S_\lambda, C_i >), prob(Label\,|{<}\,RS_\lambda, AC_{ij} >\}$$

Observations and Feature Functions

Given the training dataset, Ґ, what is known about it? In other words, what kind of observations, the feature functions need to take into consideration. In the following, we list some of the observations. First, we explain some notation. Let *Sim(X)* denote the set of similar lexicon of a lexical entity denoted by *X* and let P^{-1} denote the passive voice of a predicate denoted by *P*.

1. Two sentences with no common required lexicon or similar lexicon are unlikely to match[5] e.g. "Jury decided John is guilty" does not entail "Trees bud in spring".
2. A sequence of lexicon including morphology in one sentence matches the exact sequence of lexicon including morphology in another sentence e.g. "George saw flowers on the trees" entails "George sees flowers on the tree".
3. A predicate, *P*, with subject *S* and object *O* in one sentence matches a predicate, *P*, or one of its similar lexicon *Sim(P)* with a subject *S* or *S'* where $S' \in Sim(S)$ and an object *O* or *O'* where $O' \in Sim(O)$ e.g. "The court sentenced the criminal" entails "the jury sentenced the murderer".
4. A predicate, *P*, with subject *S* and object *O* in one sentence matches P^{-1} with subject *O* or *O'* where $O' \in Sim(O)$ and object *S* or $S' \in Sim(S)$ e.g. "the court sentenced the criminal" entails "the murderer was sentenced by the court".
5. A negated role does not match a positive role e.g. "does not entail" does not match "entails".

[4] In reality, this is not necessarily true.

[5] Note that it is not impossible that a similar lexicon is replaced by a gloss, a definition, or an inference.

6. A past participle (*VBN_verb*) could be used as an adjective; hence, its similar lexicon could be adjectives e.g. “the animal is infected” entails “the animal is sick”.
7. Complement of an auxiliary (a noun phrase, an adjective, or a prepositional phrase) could be replaced with another complement e.g. “the animal is hospitalized” entails “the animal is in the hospital”.
8. Ergative verbs need special rules: when ergative verbs have a subject but no object, we consider the subject as the object in our matching. For example, “The pollution decreased the fish populations.” and “The fish populations decreased.”'
9. A relative pronoun could be replaced with its corresponding role in the independent clause that the relative clause depends on e.g. “The Alameda Central which is west of the Zocalo was created in 1592” entails “Alameda Central is west of the Zocalo”.
10. Students may write an interrogative utterance for a statement and still be considered correct e.g. “how is rain formed?” entails “rain is formed by seeding”.

To date, our observations are represented in terms of 5 different types of feature functions:

► **Numerical or Counts:** This type captures a count of the occurrence of expected lexical entities regardless of position. A “position” does not necessarily mean an occurrence in a particular position in the sentence but an occurrence in a particular functional role in a clause or a syntactic unit. An example of this type is “NMW” in Table 1.

► **Existential:** This type captures the expected occurrence in a response sentence of a linguistic feature relating to a lexical entity, a clause, or a group of lexical entities but does not necessarily form a clause or any particular unit and the position it occurs in. An example of this type is “argsMismatch”. Basically, some of what is expected exists but not necessarily in the “order” expected.

Table 1. Examples of feature functions

Function	**Type**	**Description**
NMW	Numerical	Number of missing required entities
argsMismatch	Existential	Required argument does not match
argRoleIncompatible	Universal	Found term(s) of required argument, but the type of matching role is incompatible with type of required argument role
VPPOSMismatch	Swappers	Adjectives can be classified as *VBN_verbs* and vice-versa but other roles might not be accepted
VPPolarityMismatch	Polarity	Required role and matching role do not agree on negation

► **Universal:** This type captures the occurrence of lexical entities and corresponding linguistic features and position in a response sentence. An example of this type is argRoleIncompatible in Table 1. In this case, all that is expected exists but not necessarily in the “order” expected.

► **Swappers:** This type captures the occurrence of confusable or similar linguistic phenomenon. An example is VPPOSMismatch in Table 1.

► **Polarity:** This type captures the polarity of a linguistic constituent i.e., a word or a group of words that functions as a single unit within a hierarchical structure – to date we only consider syntactic structures. An example is VPPolarityMismatch in Table 1.

Putting it all together

Given a student response and a concept, the two texts are processed linguistically. A text is first processed for **spelling corrections** in an attempt to decrease the noise for subsequent NLP tools. Next, **parts-of-speech tagging** and **parsing** are performed. In the third stage, a parse tree is passed through a **feature extractor**. Manually-generated rules extract features from the parse tree. The result is a flat structure representing phrases, predicates, and relationships between predicates and entities. Each phrase is annotated with a label indicating whether it is independent or dependent. Each entity is annotated with a syntactic and semantic role. Then, pronouns are resolved either to an entity in the student's response or the question. Finally, a **morphology analyzer** reduces words in the mentioned flat structure to their lemmas.[6] In addition to these linguistic features, we use a manually-built thesaurus (WordNet http://wordnet.princeton.edu/) and an automatically-generated "thesaurus" (Dekang Lin's database (cs.ualberta.ca/~lindek/downloads.htm)) in order to automatically obtain a list of similar words to lexical entities. An example showing a student's response compared to a concept is shown in Table 2 including a linguistic analysis. The feature functions obtained are: nmw=0, argsMismatch:subj, argsMismatch:obj, argRoleIncompatible:subj → psubj and argRoleIncompatible: obj → pagent.

The value of the *argRoleIncompatible* function captures the observation we made about passives earlier: instead of the term being a subject (of an active verb) it was found as a psubj or subject of a passive. MaxEnt learner outputs a probability of 0.36 for *prob (1|<”the author wis seen by therobin”, “The author saw a robin”>).*

TABLE 2. An example showing a student's response compared to a concept

CONCEPT: The author saw a robin. REQUIRED WORDS: saw, robins (no similar words for robin, similar words for 'saw'={detect, perceive, notice, discover, find, observe, understand, realize}) PREPROCESS: The author saw a robin. PARSE: (TOP (S (NP (DT the) (NN author))(VP (VBD saw) (NP (DT a) (NN robin)))(. .))) LINGUISTIC ANALYSIS: Independent_clause saw :subj author :obj robin	STUDENT'S RESPONSE: the author wis seen by therobin. PREPROCESS: the author was seen by the robin. PARSE: (TOP (S (NP (DT the) (NN author)) (VP (VBD was)(VP (VBN seen) (PP (IN by)(NP (DT the) (NN robin)))))(. .))) LINGUISTIC ANALYSIS: Independent_clause be seen :psubj author :by :pagent robin

In the evaluation below, a probability ≥ 0.5 is used to decide a *Label* of 1.

[6] We do not go into detail, assuming that the reader is familiar with the described NLP techniques. In this paper, we use "term," "word," and "lexical entity" interchangeably to mean either unigrams or bi-grams.

EVALUATION

From a set of 7th- and 8th-grade items, eight Reading Comprehension (RC) and ten Maths items were selected for this experiment. The short textual responses were collected in schools in Maine, and Massachusetts, USA. Score points range from 0 to 3 and the number of concepts for each item range from 1 to 7 concepts. Table 3 shows the results for the RC items and Table 4 for the Maths items. Scoring agreement is reported in terms of quadratic kappa statistics. “H1-H2” denotes the scoring agreement between the two humans, (*c-H1/H2 MaxEnt:C*) denotes the average scoring agreement between c-rater and each human given the concept only. (*c-H1/H2 MaxEnt:C+A*) denotes the agreement with an average number of 7 alternatives provided with each concept. “Blind” denotes the set of unseen students’ responses for which MaxEnt model is used to decide an entailment. The column “*c-H1/H2 Rule:C+A”* denotes the results if we were to use a rule-based pattern matching that gives a 0/1 match (not a probability) and where observations were written in terms of rules.

TABLE 3. c-rater’s performance for the Reading items

Item	**Score points**	**Blind**	**H1-H2**	**c-H1/H2**	**c-H1/H2**	**c-H1/H2**
				Rule:C+A	**MaxEnt: C**	**MaxEnt:C+A**
R1	{0,1}	52	0.41	0.24	0.44	0.48
R2	{0,1}	54	0.68	0.08	0.63	0.71
R3	{0,1}	51	0.87	0.00	0.59	0.76
R4	{0,1}	53	0.92	0.00	0.46	0.75
R5	{0,1}	50	0.86	0.00	0.62	0.62
R6	{0,1,2}	114	1.0	0.97	0.50	0.98
R7	{0,1}	113	0.76	0.59	0.73	0.72
R8	{0,1,2}	107	0.99	0.93	0.40	0.96

MaxEnt classifier achieved much higher accuracy on most items in both RC and Maths items as the tables show though the model is built using a tiny size of randomly selected training dataset consisting of potentially very 1) noisy sentences and 2) sparse data. The worst performance, given a concept with no alternatives, was for M5. In general, having additional alternatives gave better results except in case of M8 and M9.

TABLE 4. c-rater’s performance for the Mathematics items

Item	**Score points**	**Blind**	**H1-H2**	**c-H1/H2**	**c-H1/H2**	**c-H1/H2**
				Rule:C+A	**MaxEnt:C**	**MaxEnt:C+A**
M1	{0,1,2}	96	0.97	0.01	0.15	0.76
M2	{0,1,2}	95	0.90	0.44	0.30	0.68
M3	{0,1,2}	50	0.87	0.00	0.55	0.83
M4	{0,1,2,3}	96	0.93	0.65	0.27	0.64
M5	{0,1,2}	75	0.70	0.06	0.03	0.52
M6	{0,1,2,3}	71	0.86	0.40	0.47	0.71
M7	{0,1,2}	51	0.91	0.29	0.15	0.60
M8	{0,1,2}	61	0.79	0.00	0.31	0.27
M9	{0,1,2}	49	0.46	0.00	0.59	0.56
M10	{0,1,2}	132	0.71	0.61	0.10	0.61

CONCLUSION

We have shown that a MaxEnt classifier trained on a small sample of noisy data achieves promising results for the TE task. The model achieves higher accuracy than a rule-based approach for the task of automatic content scoring. We use a MaxEnt classifier because: (1) the linguistic-based functions are overlapping and MaxEnt allows for a large number of additional functions that will further increase performance, (2) in previous NLP applications MaxEnt has achieved state-of-the-art accuracies with less supervision compared to other kind of classifiers, and (3) using the OpenNLP MaxEnt package, we needed to only provide the set of classes, the set of possible textual material, and the set of feature functions. That said, we are currently investigating the comparison of MaxEnt performance to the one of other classifiers trained on the same dataset. In the future, we would like to a) increase the set of training data, b) investigate using a training set consisting of text that is not restricted to sentences, c) investigate learning the feature functions automatically and d) calculate $prob(Label\,|< S\,\mathrm{R}, C_i >)$ by considering functions other than maximum.

ACKNOWLEDGMENTS

For Tom Morton who started the use of MaxEnt Modeling in the NLP group at ETS.

REFERENCES

Berger, A. L., Pietra S. P., & Pietra, V. J. D. (1996) A Maximum Entropy Approach to Natural Language Processing. Computatinal Linguistics 22(1) : 39-71.

Burstein J., Marcu D. & Knight, K. (2003). Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. IEEE Intelligent Systems, pp. 32-39.

Chieu, H. L., & Ng, H. T (2002). A Maximum Entropy Approach to Information Extraction from Semi-Structured and Free Text. In Proceedings of AAAI.

Christie, J. R. (2003). Automated essay marking for content- does it work ? In Proceedings of the 7th International Computer Assisted Assessment Conference. Loughborough University. Loughborough.

Goldwater, S., and Johnson, M. 2003. Learning ot constraintrankings using a maximum entropy model. In Stockholm Workshop on Variation within Optimality Theory.

Mikheev, A. (1999). Feature Lattices and maximum entropy models. Machine Learning. Proceedings of the 17th international conference on computational linguistics. Montreal.

Ming, Y. Mikhailov A., & Kuan T. L. (2000). Intelligent Essay Marking System. Learners Together, NgeeANN Polytechnic, Singapore.

Nigam, K., Lafferty, J. & McCallum, A. (1999). Using Maximum Entropy for Text Classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pp. 61-67.

Osborne M. (2002). Using maximum entropy for sentence extraction. In *Proceedings of the ACL 2002 Workshop on Automatic Summarization*,Philadelphia.

Ratnaparkhi, A. (1997). A simple introduction to maximum entropy models for natural language processing. Technical report, University of Pennsylvania.

Rehder, B. Schreiner, M. E. Wolfe, B. W., Laham, D, Landauer, T. K., & Kintsch, W. (1998). Using Latent Semantics Analysis to assess knowledge. Discourse Processes, 25, 337-354.

Rudner, L. M., & Liang, T. (2002). Austomated Essay Scoring Using Bayes' Theorem. In Proceedings of the annual meeting of the National Council on Measurement in Education.

Valenti, S, Neri F. and Cucchiarelli, A.. (2003). An Overview of Current Research on Automated Essay Grading. Journal of Information Technology Education. Volume 2.