SYMBOLIC ASSESSMENT OF FREE TEXT ANSWERS IN A SECOND-LANGUAGE TUTORING SYSTEM

Matthieu Hermet and Stan Szpakowicz

Symbolic Assessment of Free Text Answers in a Second-Language Tutoring System

Matthieu Hermet *
Stan Szpakowicz *†
* School of Information Technology and Engineering
University of Ottawa
Ottawa
Canada
† Institute of Computer Science
Polish Academy of Sciences
Warsaw
Poland
mhermet,szpak@site.uottawa.ca

Abstract

We present an approach to Computer-Assisted Assessment of free-text material based on symbolic analysis of student input. The theory that underlies this approach arises from previous work on DidaLect, a tutoring system for second-language reading skill enhancement. The theory enables the processing of free-text segments for assessment to operate without pre-encoded reference material. A study based on a corpus of 48 student answers to several types of questions has justified our approach, helped define a methodology and design a prototype.

Preliminaries

In the field of Computer-Assisted Assessment (CAA), automated processing of free-text material received from students is becoming a necessity. The range of such material may run from single sentences to whole essays. Even as seemingly small a problem as student answers to open-ended questions poses a variety of serious Natural Language Processing (NLP) challenges. It calls for different approaches, depending on the didactic purpose of the exercises. This, in turn, affects the nature of the textual material that can be submitted to automated assessment.

In NLP, there is a conceptual opposition between symbolic and statistical processing. While the first relies on methods of qualitative analysis, the second uses the distribution of quantitative text features to draw conclusions. The latter is unquestionably powerful when annotated reference material is available. This is what the field of Machine Learning calls *training data*, while the actual student material is referred to as *test data*. Assessment based on statistical technologies would mean finding the closest possible match

between the training and test material based on features. Feedback associated with the found reference match—the assessment—would then be sent to the user, in the form of a mark or comments. A major drawback of this approach is the need to have annotated reference material. It usually means a considerable amount of time and effort. Statistical methods are also by definition inaccurate, even if accuracy of over 90% is not uncommon in some language processing tasks.

Symbolic processing, on the other hand, usually relies on hand-crafted rules of analysis. It is not necessary to annotate large amounts of reference material, though crafting the right rules also takes time. Rules are triggered by feature values which tend to be acquired automatically. Performance may suffer if feature value acquisition is burdened with error. Still, it is fair to say that the very nature of the didactic process and natural languages (especially the number of exceptions at the lexical and semantic level) make exact rules preferred to nearly exact statistical methods.

Ideally, a hybrid approach—collaboration between symbolic and statistical methods—would be the best for the successful future of NLP. This is by and large a matter of NLP research, external to the concerns of CAA.

In CAA, statistical, or quantitative, processing has been preferred for, as it seems, two main reasons. The first is the existence of vast amounts of (passed) student essays or completed drills. This *is* a rich archive of problems already solved. The second reason has to do with applicability: coupled with dialogue, authoring and moderation modules, such CAA tools are reliable and work predictably well. The level of performance depends mainly on the volume of annotated material. Such systems make good summative assessment tools due to their good capacity to recognize correctness within well-defined domains.

The distinction between summative and formative assessment is not always clear. If we are to treat them as opposed to each other—a means to enhance skills through qualitative evaluation versus a means to judge skills through quantitative evaluation—building ensembles of annotated corpora rich enough to enable fully informative feedback can become a vast problem. That is because it would imply annotating all answer possibilities, including (potentially unlimited) incorrect material.

This is a rough view, and again, in practice existing systems tend to exhibit a mixture of both approaches. We believe, however, that our considerations raise the question of finding or using symbolic methods to cope with free-text analysis. Conversely, if we are to understand the problem as one of economy of annotated reference material, the question is this: is there a point in the relation between answer expressiveness and the nature of exercises, beyond which no pre-encoded answers are needed to properly perform assessment?

This is where the interest of our project lies. It originated in another project, *DidaLect*, with its strong foundation of theory of second language learning.

There is a trade-off in CALL in general between the need to design generic solutions to enhance the visibility on the marketplace (SCORM [1]) and the need to keep the tools very specific in order to guarantee reliability (Chen *et al*. [5])—this extends to CAA. Our own interest is in specificity for the sake of

demonstration: to find a proper didactic niche to implement successful symbolic solutions to prove the soundness of symbolic free-text processing within CAA or, more modestly, to test its feasibility.

The Problem

DidaLect (Balcom *et al*. [4], Desrochers *et al*. [7]) is an adaptive didactic software designed to enhance the reading ability of French-as-a-Second-Language (FSL) students working autonomously. It is firmly rooted in theories coming from the fields of education, cognition and psycholinguistics. Its Virtual Learning Environment is composed of a placement test, a tutorial and resources which support the acquisition of reading skills, for example dictionaries. *DidaLect* is therefore a good example of so-called eLearning Intelligent Tutoring System. First, the Computer Adaptive Placement Test (CAPT) (Laurier [14]) evaluates the learner on her level of French. Next, the learner is directed to a series of texts of varying difficulty, coupled with a set of comprehension-testing multiple-choice questions. The system selects text difficulty as a function of the CAPT results and the test results for the current text.

The theory behind *DidaLect*'s implementation is of crucial importance to the basic design of our free-text answer processing module, which strongly delimits the nature of questions that the student can be asked. We believe that placing such limitations on question types, assuming a solid theoretical foundation, is half of the job of building an unsupervised free-text CAA module. Very briefly, an important aspect of text comprehension is to understand the communication goals expressed by means of language. Such goals are accessible through cognitive operations of sense acquisition as well as through the awareness one has of these operations. All this is embedded in the common cultural background of the author and the reader (Duquette *et al*. [8]).

Assessment

Our system, yet unnamed, is not intended to mark answers, but rather to provide evaluation to the user on the quality of their material, in linguistic terms and on content in relation to the reference. No matter how good a CAA system is, no such system can cope with so-called bad-faith user material, such as answers correctly formulated, but deliberately crafted to fool the machine. Ellipsis, for instance, is a fine rhetorical way to answer a question, but no system can get its accuracy. So, the role of the lecturer is merely to create questions, which only requires knowledge of question categories in the field of text comprehension.

There are a number of implemented open-text CAA systems, often commercial, such as E-rater [3] and Qualrus [10]. E-rater is an Automated Essay Scoring (AES) system, marking and evaluating essays based on a set of pre-scored essays. Human raters mark training-set essays on content and fluency through the evaluation of variables, to be correlated automatically by the system in order to grant a mark. In real-world situation, E-rater is used in combination with human raters to properly assess essays. Qualrus is presented as an "Intelligent Qualitative Analysis Program". It functions as a toolbox for designing assignments as well as assessment tasks. Its

assessment capabilities are a function of both integrated NLP tools and lecturer encoding of what is to be assigned. This makes it an authoring tool rather than a straight CAA module, but it nevertheless can perform tasks of open-ended question marking and evaluation.

Texts

According to literature on the subject, there are two main types of texts: narrative and informative (Chiasson [6]). Informative texts are supposed to exhibit more complex and varying structure, which makes them more difficult to comprehend; on the other hand, they lend themselves more easily to categorization. All texts in the present prototype of *DidaLect* are informative texts, divided between four categories with fairly balanced membership: description, comparison, cause-effect and problem-solution (Richgels *et al*. [11]). The texts are news articles from general or popular-science publications. A text has normally 1-2 pages.

Questions

The categorization of questions works along two dimensions: the cognition processes needed to build understanding, and the form. Cognitively, there are three main categories of questions, addressing three forms of comprehension: literal, interpretative and critical (Chiasson [6]). It is quite difficult (or perhaps not yet feasible) to automate assessment processes for open-ended answers to questions in the two last categories. We can only realistically deal with literal comprehension questions, which have to do mainly with definitions and causal relations in texts.

Categorization by form recognizes Text-Explicit, Text-Implicit and Script-Implicit questions (Pearson *et al*. [9]). The last of these categories requires that the learners perform inference between the text and their own world knowledge; this makes answers in this category difficult to process automatically. The other two categories allow answer construction by recovering (maybe partially) the necessary fragments from one or a few sentences in the text.

If we retain only the first cognition category and the two first form categories, we believe that the resulting questions lead to open-ended answers which can lend themselves to automatic assessment processing.

- Text Explicit questions: dependence on a single sentence

> […] Comme l'avaient calculé les astronomes, l'année tibétaine 1999 débute le 16 février, lors de la nouvelle lune. **Certaines années, pour contourner des conjonctions planétaires de mauvais augure, les Tibétains suppriment des mois du calendrier ou en ajoutent d'autres**. Dans ce cas, la période du Nouvel An, appelée Lhossar, peut tomber un mois avant ou après, par rapport à notre calendrier occidental. […]

Q: Pourquoi les tibétains suppriment ou ajoutent-ils certains mois au calendrier?

- Text Implicit questions: dependence on several sentences, adjacent or (rarely) dispersed in the text.

In the following example, the sentences are not co-referenced. In such cases, we choose to encode question in two ways, one to be displayed and one to be kept by the system in a “closure” form (“replace *quoi* by the answer”).

> […] Abraham, lui, avait compris qu'il fallait sacrifier son fils à son dieu. Quelle bêtise, dirions-nous aujourd’hui! **Vouloir sacrifier son fils à son dieu. Il faut vraiment être primitif. Et pourtant, je me demande si les sociétés modernes, y compris notre société québécoise, ne sont pas un nouvel Abraham qui sacrifie de nouveaux Isaac à quelques divinités. […]**

Q_display: Comment l’auteur juge-t-il l’infanticide sacrificiel?

Q_machine: Il faut vraiment être quoi pour vouloir sacrifier son fils à dieu?

In a more complex case, the sentences are co-referenced, which enables dynamic tracking of the reference sentences making the answer using co-reference resolution techniques.

> […] Quand survient l'impact, on assiste à une réaction en chaîne: **le détecteur de décélération situé à l'avant du véhicule génère instantanément un courant électrique, qui déclenche une amorce, qui elle-même enflamme un mélange allumeur. Ce dernier met finalement le feu à l'agent propulseur responsable du gonflement du coussin**. Toute l'opération se déroule extrêmement rapidement, soit à 300 km/h. […]

Q: Quelle est la réaction en chaîne qui se produit lorsque survient un impact?

We consider that it is possible to address the issue of assessing free-text answers for such types of questions as long as the original text is known to the system.

Processing

It is a two-phase procedure to automate the assessment of free-text answers to the types of questions such as those presented in the preceding section. The first phase checks the content. It consists in comparing the learner answer LRN with the reference answer REF, represented by the text segment from which the question has been built. The second phase checks the syntactic and lexical form. Actually, the two steps are combined in the sense that content assessment works on the results of form analysis. This design seems logical, because lexical selection shapes the content as much as it affects the syntactic form.

Briefly, the procedure proceeds as follows:

1. Create words lists:
    a. words of LRN absent in REF,
    b. words of REF absent in LRN,
    c. words uncommon.
2. Perform dependency parsing of LRN and REF, producing certain dependency relations among lexical items.
3. Use a dictionary of synonym to identify synonymy between words on lists 1a and 1b.

4. Use the dependency relations from step 2, beginning with those containing synonyms found in step 3, to build trees for both sentences. Building is done by breadth-first search, which maximises the probability of discovering new/different lexical material.
5. When the process halts, trees should be completed, as should be records of any diverging lexical material between LRN to REF.
6. Check the syntax of LRN to verify if it conformity to REF, either by
    a. identity: LRN and REF have same structure,
    b. equivalence: sentence LRN is a syntactic equivalent of sentence REF, using certain pre-encoded equivalence rules.

This procedure allows us to capture student errors as follows:

1. agreement: step 2,
2. orthography: step 2,
3. synonyms: step 3,
4. missing content: step 4,
5. syntax in general: step 5.

This procedure does not yet cope with the evaluation of supplementary material. The problem is that of computing the value (in terms of contents compared with REF) of any kind of supplementary material which a student can put in the answer. At present, we can address this issue only partially by comparing the supplementary segment with the rest of the text from which REF comes. This can be explained by our observation that students tend to mix various parts of the text in their answers. Then, we can use co-reference to judge to some degree the coherence of the addition. This further procedure amounts to answering the following question: does the supplementary material interact with the theme of the question somewhere in text? And if it does, at which syntactic level? This is, however, a somewhat uninformed way of solving the problem, without regard to deep semantics. It is a partial solution which has not been tested yet.

Example

> « Selon Yves Grimard et Serge Tremblay, les précipitations acides agissent sur les écosystèmes lacustres depuis 75 ans, **soit depuis l'essor de l'industrialisation et du transport automobile**. *Au cours du XXe siècle, l'acidité des lacs de l'Outaouais s'est multiplié par 10 environ*, ce qui est trop rapide pour qu'un organisme vivant s'y acclimate. »

The following question is Text-Implicit. In order to link the two sentences needed to relate question and answer fragment (*Italics* and **bold**), the question is also encoded under closure form.

> Q_display: Pourquoi l'acidité des lacs de l'Outaouais s'est-elle multipliée par 10 au cours du XXème siècle?

Q_machine: Depuis quoi les precipitations agissant sur les ecosystèmes lacustres ont multiplié par 10 l'acidité des lacs de l'Outaouais*?*

S1: Depuis l'essor de l'industrialisation et du transport automobile[1].

S2: A cause de l'essor de l'industrialisation et du transport automobile.

Creating word lists for S1 will signal the identity of form, as lists 1 and 2 are empty. The list of words in common contains all words of both chunks REF and S1. In such cases, a mere surface comparison of REF and S1 will suffice to assess S1.

Creating words lists for S2 will yield the following result:

- L1: A, cause, de
- L2: depuis
- L3 ; essor, industrialisation, transport, automobile[2]

There is no synonymy relation between *cause* and *depuis*. But checking *cause* in the synonymy dictionary will enable detection of the compositional form of *à cause de*.

A fourth list is created to record words present in Q_display and absent from the set of words contained in both Q_machine and answer segment. This only yields *pourquoi* which is synonymous with *cause*, as shown by Memodata [2]. We have no means of knowing whether *cause* stands for *depuis*, but at this stage we know that it correctly corresponds to the question marker *pourquoi*. As the system cannot go any further in lexical comparison, it moves to the next step, parsing.

S2, (partial) syntactic analysis using XIP [12]

```
NMOD_POSIT1_RIGHT_ADJ(transport,automobile)
NARG_POSIT1_CLOSED_NOUN_INDIR(essor,de,industrialisation)
COORDITEMS_CLOSED_PREP_NOUN(essor,transport)
PREPOBJ_CLOSED(A cause de,essor)
PREPOBJ_CLOSED(de,industrialisation)
PREPOBJ(du,transport)
0>GROUPE{PP{A cause de NP{l' essor}} PP{de NP{l' industrialisation}} et PP{du NP{transport}} AP{automobile} .}
```

REF

```
NMOD_POSIT1_RIGHT_ADJ(transport,automobile)
NARG_POSIT1_CLOSED_NOUN_INDIR(essor,de,industrialisation)
COORDITEMS_CLOSED_PREP_NOUN(essor,transport)
PREPOBJ_CLOSED(Depuis,essor)
PREPOBJ_CLOSED(de,industrialisation)
```

[1] These are the two answers we obtained for the question. These should show the tendency of students to re-use text chunks.

[2] Function words are discarded from L3.

PREPOBJ(du,transport)

1>GROUPE{PP{Depuis NP{l' essor}} PP{de NP{l' industrialisation}} et PP{du NP{transport}} AP{automobile} .}

As we cannot initiate tree-building starting with synonyms (there are none) and as there is no verb phrase to choose as sentence head, the order is to begin with the first relation in the analysis[3]. Here, it is the same for both:

NMOD_POSIT1_RIGHT_ADJ(transport,automobile)

Tree-building performs as follows.

- Retrieve all relations in which a modified term appears (here, *transport*):

COORDITEMS_CLOSED_PREP_NOUN(essor,transport)

- Merge the relations:

COORDITEMS_CLOSED_PREP_NOUN(essor,
NMOD_POSIT1_RIGHT_ADJ(transport,automobile))

This composite relation here is the same for both sentences. This determines the selection of a word on which to iterate merging. The policy is to select the most promising word in terms of semantic importance, or in terms of the probability of discovering supplementary material. To simplify, the resulting complete composite relations are as follows, getting rid of DET relations:

COORDITEMS_CLOSED_PREP_NOUN(PREPOBJ_CLOSED(Depuis,
NARG_POSIT1_CLOSED_NOUN_INDIR(essor,de,industrialisation)),
NMOD_POSIT1_RIGHT_ADJ(transport,automobile))

COORDITEMS_CLOSED_PREP_NOUN(PREPOBJ_CLOSED(A cause de,
NARG_POSIT1_CLOSED_NOUN_INDIR(essor,de,industrialisation)),
NMOD_POSIT1_RIGHT_ADJ(transport,automobile))

Two conclusions can be drawn from this process and analysis. First, *à cause de* as well as *depuis* have both been recognized at parsing time as prepositional phrase heads. Second, the student neither added nor subtracted any textual material with respect to the reference answer. We know, therefore, that no lexeme has undergone any reformulation and that the sentences have identical syntactic structure. As *Á cause de* has also been recognized as a proper answer connector to why-questions, and as it fits the sentence syntax, S2 will be assessed as correct.

Assessment

The example we followed in section 3 shows no errors. We chose it to keep the explanation short while still describing the processing possibilities. The errors, if any, are captured during processing. We examine in turn all types of errors.

[3] The policy for the selection of relations, in case the system has to choose between several, is to favour higher-order categories (SUBJ, OBJ, REL, COORDITEMS…) over lower-order (NMOD, ADJMOD, PREPOBJ…).

Ortography and Agreement

XIP (Aït-Mokthar *et al*. [13], [12]), our parser, outputs the number and gender of the words in addition to what has been shown. A comparison between the lexical files of LRN and REF is all we need to assess the contents with respect to orthography and agreement. This poses the question of number generalization (*les hommes* can be equivalent to *l'homme*), as a student can choose to use singular for plural in an attempt to generalize number. This problem has been left for future work.

Synonymy

The system can only give a partial judgement on the exact pertinence of lexical reformulation. Synonymy is easy to detect with *Memodata basis*, the synonymy dictionary [2], even across parts of speech. Errors are simply recorded as wrong lexical reformulation choices at given syntactic positions, in comparison to REF. We have no means of evaluating such errors in supplementary material. Errors in prepositions are also recorded at this stage, still using *Memodata basis*.

Content

Once the content correspondence between REF and LRN has been established when building trees, the problem is to know whether LRN contains part or all of REF, or even more than REF. Partial correspondence is detected by modifier or complement gaps in LRN with respect to REF, and can only be signalled to the user. An answer is still considered acceptable if it contains only lexical heads. Supplementary material is evaluated through syntax and through the relation which supplementary elements have to other occurrences of heads in the rest of the text.

Syntax

Syntax is assessed through rules of reformulations as well as through heuristics. Rules of reformulations establish correspondence between structures equivalent in meaning but different in form. Those categories of reformulations include mainly nominalization, passive/active and pronominalization. This is achieved by comparing the structure of LRN and REF. Heuristics detect clause reduction in a procedure supported by lists of attribute, state and action verbs; in clause reduction, a phrase containing a verb or modified nouns is reduced to one of its member. The main idea behind this machinery is that reformulation has recursive power: it can occur at the level of the whole sentence or at the phrase level.

Future Work and Conclusion

To keep the list of future tasks short, we prefer future work to strengthen what has already been achieved rather than adding functionality. That is why our main objective is to have an exhaustive set of reformulation rules and heuristics in order to address typical mistakes that FSL students commit, as observed in a set of fifty 20-page journals written by FSL students.

In the present state, we can recognize 46 answers (to 16 questions) out of 48 answers gathered from students during experiments.

Acknowledgement

This work has been partially supported by the Social Sciences and Humanities Research Council of Canada, in the program "Initiative on the New Economy".

References

[1] SCORM/ADL (Advanced Distributed Learning) at http://www.adlnet.org

[2] Alexandria by Memodata at http://www.memodata.com

[3] Y. Attali, J. Burstein (2006). "Automated Essay Scoring with e-rater® V.2.". *Journal of Technology, Learning and Assessment, 4(3)*. Available from http://www.jtla.org

[4] P. Balcom, T. Copeck, S. Szpakowicz (2006). "DidaLect: Conception, Implantation et Evaluation Initiale". Technologies Langagières et Apprentissage des Langues: Actes du Colloque tenu dans le Cadre du Congrès de l'ACFAS, 11-12 mai 2004, sous la direction de L. Duquette et C. St Jacques, Montréal: ACFAS Cahier No. 105.

[5] L. Chen, N. Tokuda (1999). "A New Diagnostic System for J-E Translation ILTS". *Proc Machine Translation Summit*, 608-616.

[6] J. Chiasson (1990). *La Compréhension en Lecture*. Gaëtan Morin Eds, Montréal.

[7] A. Desrochers, L. Duquette, S. Szpakowicz (2004). "Adaptive Courseware for Reading Comprehension in French as a Second Language: The Challenges of Multidisciplinary in CALL". *Proc 11th international CALL Conference*, Antwerpen, 85-91.

[8] L. Duquette, A. Desrochers (2006). "Appuyer la Compréhension en Lecture à l'Aide d'un Logiciel Adaptatif". Technologies Langagières et Apprentissage des Langues: Actes du Colloque tenu dans le Cadre du Congrès de l'ACFAS, 11-12 mai 2004, sous la direction de L. Duquette et C. St Jacques, Montréal: ACFAS Cahier No. 105.

[9] D. Pearson, D. Johnson (1978). *Teaching Reading Comprehension*. New York, Holt, Rinehart and Winston.

[10] Qualrus by Ideaworks at http://ideaworks.com/qualrus.shtml

[11] D. Richgels, L. McGee, R. Lemax, C. Sheard (1987). "Awareness of Four Text Structures: Effects on Recall of Repository Texts". *Reading Research Quarterly XXII*, 177-197.

[12] XIP Parser, Xerox Research Center Europe Technical Document, 2003.

[13] S. Ait-Mokhtar, J.-P. Chanod, C. Roux (2001). "A Multi-Input Dependency Parser". *Proc. Seventh IWPT (International Workshop on Parsing Technologies)*, Beijing.

[14] M. Laurier (1999). "The development of an adaptive test for placement in French". M. Chalboub-Deville (ed.), *Development and research in computer adaptive language testing*. Cambridge: University of Cambridge Examinations Syndicate / Cambridge University Press, 122-135.