Tracking the Evolution of Written Language Competence: an NLP–based Approach

Stefan Richter•, Andrea Cimino◇, Felice Dell'Orletta◇, Giulia Venturi◇
•University of Leipzig (Germany)
hewuri@gmail.com
◇Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{name.surname}@ilc.cnr.it

Abstract

English. In this paper, we present an NLP-based innovative approach for tracking the evolution of written language competence relying on different sets of linguistic features that predict text quality. This approach was tested on a corpus essays written by Italian L1 learners of the first and second year of the lower secondary school.

Italiano. *In questo articolo, presentiamo un metodo innovativo per monitorare l'evoluzione delle competenze di scrittura basato su tecnologie del linguaggio che sfruttano caratteristiche linguistiche predittive della qualità del testo. Questo approccio è stato testato su un corpus di produzioni scritte di apprendenti l'italiano L1 del primo e secondo anno della scuola secondaria di primo grado.*

1 Introduction and Background

Using automatic techniques to trace the learning progress of students starting from their written productions is receiving growing attention in many different research fields and for different purposes. Two different perspectives are taken into account: i.e. the analysis of the *form* and of the *content* of texts. The first scenario is mainly addressed within the writing research community where the learning progress is framed as an analysis aimed at detecting linguistic predictors of written quality across grade levels. Using Natural Language Processing (NLP) tools, different set of features (e.g. grammar features, errors, measures of lexical complexity) are automatically extracted from corpora of student essays to investigate how they relate to writing quality (Deane and Quinlan, 2010) or to other literacy processes such as reading (Deane, 2014). Human ratings of essay writing quality are also used to develop Automatic Essay Scoring systems mostly of L2 essays (Attali and Burstein, 2006). For what concerns the analysis of *content* of texts, traditional Knowledge Tracing systems are based on a framework for modeling the process of student learning while completing a sequence of assignments (Corbett and Anderson, 1994). These systems rely on a correctness value of each assignment given by a teacher. More recently, the Knowledge Tracing framework started to be explored by the Machine Learning (ML) community[1] in Adaptive E–learning scenarios. Different ML approaches have been devised to build statistical models of student knowledge over time in order to predict how students will perform on future interactions and to provide personalized feedback on learning (Piech et al., 2015; Ekanadham and Karklin, 2015).

Both the evaluation of *form* and *content* of a text share a common starting point: they imply a human 'commitment'. In the first case, it is assumed that the analyzed essays are manually scored according to the writing quality level, in the second case, the statistical models are trained on student exercises indicating whether or not the exercise was answered correctly.

In this paper, we present an innovative approach for tracking the evolution of written language competence using NLP techniques and relying on not-scored essays. Our approach focuses on the analysis of *form* but we combined for the first time the methods developed to tackle the form and content evaluation. Namely, we automatically extracted from written essays linguistic predictors of text quality that we used as features of a machine learning classifier to trace student developmental growth over the time. We tested this method on a corpus of written essays of Italian L1 learners collected in the first and second year of the lower secondary school. The use of not-scored essays is

[1] http://dsp.rice.edu/ML4Ed_ICML2015

one of the main novelty making our approach particularly suited for less resourced languages such as the Italian language, as far as corpora of L1 students are concerned.

2 Our Approach

Our approach of tracking the evolution of written language competence of L1 learners is based on the assumption that given a set of chronologically ordered essays written by the same student a document d_j should show a higher written quality level with respect to the ones written previously. Following this assumption, we consider the problem of tracking the evolution of a student as a classification task. Given two essays d_i and d_j written by the same student, we want to classify whether $t(d_j) > t(d_i)$, where $t(d_i)$ is the time in which the document d_i was written.

For this purpose, we built a classifier operating on morpho-syntactically tagged and dependency parsed essays which assigns to each pair of documents (d_i, d_j) a score expressing its probability of belonging to a given class: 1 if $t(d_j) > t(d_i)$, 0 otherwise. Given a training corpus, the classifier builds all possible pairs (d_i, d_j) of documents written by the same student. For each pair of documents (d_i, d_j), two feature vectors (V_{d_i}, V_{d_j}) are extracted. Exploiting these two vectors, $V_{d_i,d_j} = V_{d_i} - V_{d_j}$ is computed. Since many machine learning algorithms assume that input data values are in a standard range, we finally calculated V'_{d_i,d_j} obtained by scaling each component in the range $[0, 1]$. The classifier was trained and tested on the corpus described in section 3, it uses the features described in section 4 and linear Support Vector Machines (SVM) using LIBSVM (Chang and Lin, 2001) as machine learning algorithm.

3 Corpus

We relied on CItA (*Corpus Italiano di Apprendenti L1*), the first corpus of essays written by Italian L1 learners in the first and second year of lower secondary school which has been manually annotated with grammatical, orthographic and lexical errors (Barbagli et al., 2015). The corpus contains 1,352 texts written by 156 students and collected in 7 different lower secondary schools in Rome: 3 schools (77 students) are located the historical center and 4 schools (79 students) in suburbs. CItA contains two different types of essays differing with respect to the prompt, i.e the prompts assigned individually by each teacher during each school year and a prompt common to all schools that was assigned at the end of the first and second year. It is also accompanied by a questionnaire containing a set of questions referring to the student background (e.g. questions about the student family, about the native language spoken at home, etc.). This makes possible to investigate whether and to which extent some of the student background information are related to the observed language competence evolution. The main peculiarity of the corpus is its diachronic nature. Even though the contained essays were not manually scored, the covered temporal span makes CItA particularly suitable for tracking the evolution of the written language competence over the time.

4 Features

Our approach relies on multi–level linguistic features, both automatically extracted and manually annotated in CItA. A first set of features was extracted from the corpus morpho-syntactically tagged by the POS tagger described in (Dell'Orletta, 2009) and dependency-parsed by the DeSR parser using Multi-Layer Perceptron (Attardi et al., 2009). They range across different linguistic description levels and they qualify lexical and grammatical characteristics of a text. These features are typically used in studies focusing on the "form" of a text, e.g. on issues of genre, style, authorship or readability (see e.g. (Biber and Conrad, 2009; Collins-Thompson, 2014; Cimino et al., 2013; Dell'Orletta et al., 2014)). The second set of features refers to the errors manually annotated. Also these features range across different linguistic description levels.

Raw and Lexical Text Features *Sentence Length* and *Token Length*: calculated as the average number of words and characters. *Basic Italian Vocabulary rate features*: these features refer to the internal composition of the vocabulary of the text. To this end, we took as a reference resource the Basic Italian Vocabulary by De Mauro (1999), including a list of 7000 words highly familiar to native speakers of Italian. *Words Frequency class*: this feature refers to the average class frequency of all lemmas in the document. The class frequency for each lemma was computed exploiting the *2010-news-1M* corpus (Quasthoff et al., 2006), using the following function: $C_{cw} = \lfloor \log_2 \frac{freq(MFL)}{freq(CL)} \rfloor$, where MFL is the

most frequent lemma in the corpus and CL is the considered lemma. *Type/Token Ratio*: this feature refers to the ratio between the number of lexical types and the number of tokens.

Morpho-syntactic Features *Language Model probability of Part-Of-Speech unigrams*: this feature refers to the distribution of unigram Part-of-Speech. *Lexical density*: this feature refers to the ratio of content words (verbs, nouns, adjectives and adverbs) to the total number of lexical tokens in a text. *Verbal mood*: this feature refers to the distribution of verbs according to their mood.

Syntactic Features *Unconditional probability of dependency types*: this feature refers to the distribution of dependency relations. *Parse tree depth features*: this set of features captures different aspects of the parse tree depth and includes the following measures: a) the depth of the whole parse tree, calculated in terms of the longest path from the root of the dependency tree to some leaf; b) the average depth of embedded complement chains governed by a nominal head and including either prepositional complements or nominal and adjectival modifiers; c) the probability distribution of embedded complement chains by depth. *Verbal predicates features*: this set of features ranges from the number of verbal roots with respect to number of all sentence roots occurring in a text to their arity. The arity of verbal predicates is calculated as the number of instantiated dependency links sharing the same verbal head. *Subordination features*: these features include a) the distribution of subordinate vs main clauses, b) their relative ordering with respect to the main clause, c) the average depth of chains of embedded subordinate clauses and d) the probability distribution of embedded subordinate clauses chains by depth. *Length of dependency links*: the length is measured in terms of the words occurring between the syntactic head and the dependent.

Annotated Error Features These features refer to the distribution of different kinds of errors manually annotated in CItA: a) *grammatical errors*, e.g. wrong use of verbs, preposition, pronouns; b) *orthographic errors*, e.g. inaccurate double consonants (e.g: *tera* instead of *terra*, *subbito* instead of *subito*); c) *lexical errors*, i.e. misuse of terms.

5 Experiments and Discussion

The system was evaluated with a weighted 7-fold cross validation in which every fold is represented by a different school. It follows that in each experiment the test set is composed by the school documents which are not included in the corresponding training set. The accuracy for each fold is calculated in terms of F-Measures. The final score is the weighted average with respect to the number of student of each school.

Feature	Correlation
9 most correlated features used for feature selection	
Frequency class of verbs	0.212
Percentage of auxiliary verbs in first person plural	-0.168
Number of tokens	0.164
Number of sentences	0.162
Percentage of prepositional dependency relation	0.153
Percentage of auxiliary dependency relation	-0.137
Percentage of auxiliary verbs in indicative	-0.136
Type/Token ratio (first 200 tokens)	0.130
Average of characters per token	0.126
Correlation of manually annotated errors	
Grammatical errors per word	-0.103
Orthographic errors per word	-0.119
Lexical errors per word	-0.162

Table 1: Correlations between features and the chronological order of the texts

Four different sets of experiments were devised to test the performance of our system. The experiments differ with respect to the temporal span between the two compared documents used in training and test sets. In the first experiment all pairs of texts written by the same student are used as training and test set, which means that the sets contain pairs of documents with all possible temporal distances (from the minimum to the maximum distance). In the second experiment we compared only the texts written in two different years, so that at least one year occurs between the documents. In the third experiment the pairs used in the training and test sets contain the first and the penultimate text written by the same student, whereas in the last experiment the first and the last text of a student were compared. Thus the time period between the texts is the maximum possible, i.e. two years. Every experiment was performed using all features described in section 4 and using only a subset of features resulting from the feature selection process. These features were selected by calculating the correlation between all features (with the exclusion of the *Annotated Error features*) and the chronological order of the texts of each student. For these experiments we selected the nine

	F-Score for each school							Weighted average F-Score
	1	2	3	4	5	6	7	
all texts								
All Features	73.0	68.0	56.5	59.1	64.8	51.8	64.0	**62.7**
Feature Selection	67.3	70.9	50.2	71.4	55.9	57.4	59.5	**61.2**
Feature Selection + Errors	67.3	70.5	54.5	73.4	56.2	57.5	59.2	**61.6**
different years								
All Features	78.1	70.5	52.3	68.5	68.0	44.3	76.7	**64.1**
Feature Selection	77.9	77.4	48.4	67.5	63.6	57.5	59.1	**64.8**
Feature Selection + Errors	77.4	78.2	50.2	67.7	63.6	57.5	58.5	**64.6**
first and penultimate text								
All Features	84.0	92.6	73.9	61.9	55.6	56.5	64.3	**71.7**
Feature Selection	92.0	96.3	65.2	95.2	72.2	58.7	71.4	**79.8**
Feature Selection + Errors	92.0	96.3	70.2	96.3	72.8	62.4	71.4	**81.2**
first and last text								
All Features	100.0	96.3	87.0	81.8	76.3	95.8	78.6	**89.3**
Feature Selection	76.0	96.3	52.2	90.9	78.9	100.0	82.1	**82.8**
Feature Selection + Errors	78.2	96.3	55.2	89.7	80.7	100.0	82.3	**84.1**

Table 2: Results of experiments.

most correlated features corresponding to different linguistic phenomena, reported in Table 1.

The results of all experiments are shown in Table 2. As a general remark, we can note that the bigger the temporal span between the tested documents, the bigger the achieved accuracy. This is due to the fact that the growth of the student writing quality is related to the temporal span. The best accuracy is achieved in the *first and last text* experiment (89.2%) using all features. Since the last text is the Common Prompt written at the end of the second year, this result can be biased by the features capturing prompt-dependent characteristics rather than the language competence evolution. Therefore the result could indicate an overfitting of the model. This assumption is supported by the accuracy achieved in the *first and penultimate text* experiment using all features (71.7%). In this case, the prompts of the written essays differ from school to school.

The *Feature Selection* rows report the results obtained after the feature selection process. Even though in these experiments we considered only nine features (vs. the total number of about 150 features), we can note a general improvement in particular for what concerns the *first and penultimate text* experiment (about 8% points of improvement). These results demonstrate that these nine features are able to capture the evolution of the written language competence at different level of linguistic description. The main competence improvement captured by these features refer to: the use of verbs, in terms of both the frequency class of used verbs (during the language competence evolution the students tend to use less frequent verbs) and the verb structures produced by the students, as it is suggested by the occurrence of features capturing the use of the auxiliary verbs; basic characteristics of the sentence, such as the sentence and word length; and features referring to lexical richness (the type/token ratio feature). Interestingly, these features are in line with the results obtained by socio-pedagogical studies reported in (Barbagli et al., 2014). It is noticeable that the results of the third school are significantly the lowest ones when feature selection is used. This is due to the fact that the nine selected features do not significantly change in the student essays over the time for this school. Further investigations is part of our current studies where we are combining student background information with the competence evolution.

The *Feature Selection + Errors* rows show the results obtained using the manually annotated errors combined with the nine selected features. As we can note, in almost all cases we obtained only a small improvement with respect to the feature selection results. This result is of pivotal importance demonstrating that the written language competence is mainly captured by relying on features that refer to the essay linguistic structure rather then by focusing on errors (also when manually annotated). This is in line with the observation of De Mauro (1977) who claims that, in particular for what concerns orthographic errors, the language competence is not related with the orthography correctness.

References

Y. Attali and J. Burstein. 2006. Automated Essay Scoring with e–rater V.2. *Journal of Technology, Learning, and Assessment*, 4(3).

G. Attardi, F. Dell'Orletta, M. Simi, and J. Turian. 2009. Accurate Dependency Parsing with a Stacked Multilayer Perceptron. *Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian)*, Reggio Emilia.

A. Barbagli, P. Lucisano, F. Dell'Orletta, S. Montemagni, and G. Venturi. 2014. Tecnologie del linguaggio e monitoraggio dell'evoluzione delle abilità di scrittura nella scuola secondaria di primo grado. *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it)*, 9–10 December, Pisa, Italy.

A. Barbagli, P. Lucisano, F. Dell'Orletta, S. Montemagni, and G. Venturi. 2015. CItA: un Corpus di Produzioni Scritte di Apprendenti l'Italiano L1 Annotato con Errori. *Proceedings of the 2nd Italian Conference on Computational Linguistics (CLiC-it)*, 2–3 December, Trento, Italy.

D. Biber and S. Conrad. 2009. Genre, Register, Style. Cambridge: CUP.

A. Cimino, F. Dell'Orletta, G. Venturi, and S. Montemagni. 2013. Linguistic Profiling based on Generalpurpose Features and Native Language Identification. *Proceedings of Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, Georgia, June 13, pp. 207-215.

C. Chang and C. Lin. 2001. LIBSVM: a library for support vector machines. *Software available at* `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

K. Collins-Thompson. 2014. Computational Assessment of text readability. *Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics*, 165:2, John Benjamins Publishing Company, 97-135.

A.T. Corbett and J.R. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user–adapted interaction*, 4(4), 253–278.

P. Deane. 2014. Using Writing Process and Product Features to Assess Writing Quality and Explore How Those Features Relate to Other Literacy Tasks. *ETS Research Report Series*.

P. Deane and T. Quinlan. 2010. What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research*, 2(2), 151–177.

F. Dell'Orletta. 2009. Ensemble system for Part-of-Speech tagging. *Proceedings of Evalita'09, Evaluation of NLP and Speech Tools for Italian*, Reggio Emilia, December.

F. Dell'Orletta, M. Wieling, A. Cimino, G. Venturi, and S. Montemagni. 2014. Assessing the Readability of Sentences: Which Corpora and Features. *Proceedings of the 9th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2014)*, Baltimore, Maryland, USA.

T. De Mauro. 1977. *Scuola e linguaggio*. Editori Riuniti, Roma.

T. De Mauro. 1999. *Grande dizionario italiano dell'uso (GRADIT)*. Torino, UTET.

C. Ekanadham and Y. Karklin. 2015. T-SKIRT: Online Estimation of Student Proficiency in an Adaptive Learning System. *Proceedings of the 31st International Conference on Machine Learning*.

C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. Guibas, and J. Sohl-Dickstein. 2015. Deep Knowledge Tracing. *ArXiv e-prints:1506.05908 2015*.

U. Quasthoff, M. Richter, and C. Biemann. 2006. Corpus Portal for Search in Monolingual Corpora. *Proceedings of the fifth international Language Resources and Evaluation Conference (LREC-06)*, Genoa, Italy.