University Entrance Exam as a Guiding Test for Artificial Intelligence

Igor Cataneo Silveira
Institute of Mathematics and Statistics
University of São Paulo
Email: igorcs@ime.usp.br

Denis Deratani Mauá
Institute of Mathematics and Statistics
University of São Paulo
Email: ddm@ime.usp.br

Abstract—**In this paper, we propose using an University Entrance Exam (the Exame Nacional do Ensido Médio) as a proper test of success of artificial intelligence techniques, thus replacing the famous Turing Test. We argue that more importantly than measuring the ability of a system in replicating human competence, the so-called ENEM test can serve as a driver of development of new techniques. Additionally, we describe how we produced a machine-readable database of questions from previous exams, which can be used for comparing techniques for natural language processing, image processing and knowledge representation and reasoning. We then present some preliminaries to serve as baseline that are based on information retrieval techniques and Word2Vec. Experiments with previous exams show that in questions concerning Humanities and Languages these baseline methods perform in average slightly better than random guessing.**

1. Introduction

Since its proposal in 1950, the Turing Test has generated a strong debate over its adequacy as a measure of (artificial) intelligence [1], [2]. Criticisms towards the test range from it being overly simplistic (e.g., that a trivial symbol manipulator is enough to pass the test [3]) to it being overly complicated (e.g., that conversations are too adaptable and wide-ranging [4]).

Much less attention has been paid to the adequacy of the Turing Test as concrete driver of development in artificial intelligence (AI) research. In fact, and in spite of the occasional attempts to run competitions based on simplified versions of the test (viz. the Loebner Prize [5] and the Turing Test 2014 [6], [7]), most researchers in the field have set aside the Turing Test as a philosophical question, one that bears little influence on their practical work.

On the other hand, historically, competitions (or tests) have been a great driver of development in many fields. Take for instance the Longitude Act: created by the British Government to promote the creation of a precise, simple and practical method for determining the longitude of a ship, it spurred the development of lunar tables and the marine chronometer. Another example is the RoboCup [8], a competition held annually to "promote robotics and AI research, by offering a publicly appealing, but formidable challenge" . The RoboCup, originally a sort of soccer tournament played by teams of robots, now includes many other categories (tests) such as search and rescue missions. Progress towards its major goal (winning a match against the champion of the World Cup) has been steady and admirable: robots went from only being capable of penalty shoot-outs to passing the ball and to performing goal keeper diving [9]. As a final example, consider the self-imposed challenge taken by IBM researchers of developing a computer system that could win at the popular American TV quiz show Jeopardy!. The Jeopardy Challenge succeeded in attracting greater attention to question answering technologies, which includes, among other things, solutions for natural language parsing, (question) classification, automatic knowledge extraction, knowledge representation and automated reasoning [10].

If competitions and tests are so useful to guide scientific and technological progress, we ask: Regardless of its character as a measure of intelligence, is the Turing Test a proper driver of progress in AI? And if it is not, then what is a proper replacement?

These questions have been considered by several researchers such as Davis [11], Miyao and Kawazoe [12], to name a few. In essence, there seems to exist some consensus that a proper guide for the development of AI must (a) solve a real-world task that (b) is somehow restricted, and (c) has fair and clear evaluation, therefore (d) allowing comparison between human and machine. Clark et al. [13] added to this list the ability to perform (e) commonsense knowledge and (f) commonsense reasoning. Yet another important feature raised by Davis is to have a test that is not designed by the community that its purported to solve it (in Davis' words, to avoid "putting the fox in charge of the chicken coops").

These requirements rule out the Turing Test as a proper driver for AI research since (b) it is unrestricted, as conversations are wide-ranged; (c) it does not have a fair and clear evaluation system, as it relies on a (group of) human evaluator(s) producing a simple pass/fail output and (d) it applies only for machines — just imagine how awkward it would be to a human to fail the test.

The *Exame Nacional do Ensino Médio (ENEM)* is an academic exam widely applied every year by the Brazilian government to students that wish to undertake a University degree; as a results of this, it is considered (in part or uniquely) as entrance exam by several major universities in Brazil. The exam consists of the writing of an essay and an objective part containing 180 multiple choice questions. These questions are divided into four groups of 45 questions each: Humanities, Languages, Sciences and Mathematics. Notably, the previous exams (with solutions) are publicly and freely available (but not in a machine-readable form).

We propose the use of the ENEM as a general driver/test of development in AI research, since: (a) it is a real-world task; (b) it contains a limited, although very large, domain; (c) it consists of a sequence of multiple choice questions, and thus allows for fair and clear evaluation; (d) it allows comparison between human and machine performances through the score of both in the same exam; (e) commonsense knowledge is embedded in its questions; (f) it requires commonsense reasoning; (g) it promotes language processing techniques for (Brazilian) Portuguese.

We argue that any system that achieves human-level competence in the ENEM (measured by the number of points in the test), while possibly still very far from exhibiting (human) intelligence, must perform well in a number of useful tasks such as text and image understanding, usage of encyclopedic and commonsense knowledge, as well as the ability to perform effective (logical and probabilistic) inference. Thus, the ENEM Challenge (i.e., the idea of ranking AI systems by their score in ENEM) will both likely contribute to the advancement of the state-of-the-art in AI, and to serve as a useful indicator of the overall success of AI techniques. Much like Robocup's proponents, we believe that this challenge will promote AI research by drawing overall public interest in a scientifically worthy task.

2. Related Work

There are many proposals of replacements for the Turing Test in the literature. Levesque proposed the Winograd Scheme Challenge [4], a test based on answering binary questions about small phrases written in natural language. The phrases usually contain ambiguity that can only be resolved using commonsense reasoning. By its nature, the test requires the creation of a significant base of questions (probably by the community), and the evaluation of a system requires experimentation and repetition (to bypass the lack of objectivity in the definition of solutions).

The systems Praline [14] and Aristo [15] were developed to answer the New York Regents 4th Grade Science Test. Written in simple English, the test consists of multiple choice questions on Primary School Science topics, which generally require a great deal of commonsense and logical reasoning. Praline uses Markov Logic Networks to reason over knowledge represented in first order logic. Aristo uses a combination of several solvers that manipulate different sorts of knowledge to achieve a state-of-the-art performance with mean accuracy of 71.3% against 47.5% by Praline.

Davis argues that standardized tests are created to be hard for people, but not necessarily for computers [11]. He thus proposes creating a curated database of multiple choice science questions taken from both High and Primary School tests. The latter's questions should avail understanding: time, causality and the human body, while the former's questions should be related to scientific methodology, including interpreting real world phenomena and laboratory experiments.

The Japanese University Entrance Exam was proposed as benchmark for Natural Language Processing by Miyao and Kawazoe [12]. The test consists of questions about 10 subjects (written in Japanese) in addition to Foreign Language questions (written in English). The test was converted into a machine readable format and used in several competitions [16], [17]. Several different systems were evaluated by their capacity to recognize entailment, contradiction and independence in Japanese, English or Chinese texts. The most successful approach used a pool of different question-dependent techniques; it outperformed humans in World History, but was outperformed in other subjects.

Págico was a competition that required answering a question written in natural language (Portuguese) with a Wikipedia page containing the answer to that question [18]. Two systems participated, both used information retrieval, the first [19], transforms the question into a SPARQL query to dbpedia and retrieves information based on the answer of this database. The second [20] augmented the questions with synonyms of the identified noun and verbal phrases. The former achieved 8% accuracy, the latter, 12%.

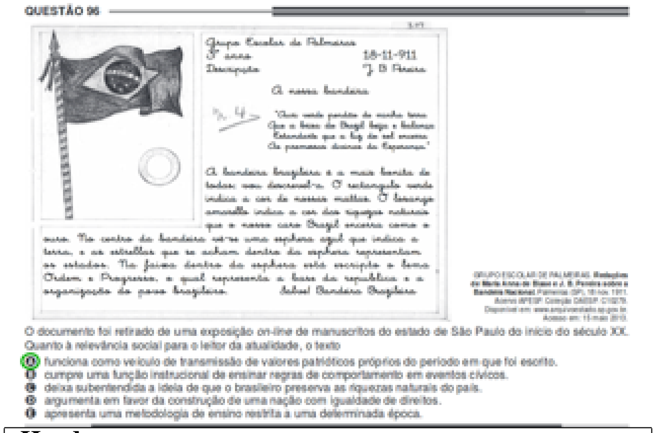
3. A Machine-Readable ENEM Database

The *Exame Nacional do Ensino Médio (ENEM)* is taken by the majority of Brazilian students who wish to enroll in a undergraduate education program. As previously stated, the objective part consists of 180 multiple choice questions evenly split into four major topics: Humanities, Languages, Sciences and Mathematics. The second part includes five questions on Foreign Language (either English or Spanish, depending on the taker's choice), which we discarded in order to simplify matters. The exam usually takes two days: Humanities and Sciences are on the first day, while Languages and Mathematics appear on the second day.

Free PDF copies of the exam's previous editions can be downloaded from its website [21]. Although each exam contains a loose pattern mixture of images and text, a question usually has the following format: a text or image is presented (the header), followed by a textual statement, then the five alternatives, one of which is correct. Sometimes, an image or text is shared by many questions.

Ultimately, we would like to evaluate an AI system by its ability on solving digital copies of the ENEM, but that might require a great deal of image processing that we fear might drive interest away. In order to maintain the focus of the test on knowledge processing, we created a machine-readable database of questions by manually converting exams into structured textual form[1] (XML format). At this initial stage, we only considered questions from Humanities and Language; we also only retained the textual part of questions and discarded any non-textual information (although we did use it to annotate questions, as will soon be explained). We segmented each question into three parts: the *header*,

[1] Available at: http://www.ime.usp.br/~ddm/enem/



QUESTÃO 96

Grupo Escolar de Palmeiras
3° anno 18-11-911
Descripção J B Pereira

O documento foi retirado de uma exposição on-line de manuscritos do estado de São Paulo do início do século XX. Quanto à relevância social para o leitor da atualidade, o texto

- funciona como veículo de transmissão de valores patrióticos próprios do período em que foi escrito.
- cumpre uma função instrucional de ensinar regras de comportamento em eventos cívicos.
- deixa subentendida a ideia de que o brasileiro preserva as riquezas naturais do país.
- argumenta em favor da construção de uma nação com igualdade de direitos.
- apresenta uma metodologia de ensino restrita a uma determinada época.

Header

Grupo Escolar de Palmeiras 3 anno 18-11-911
Descripção J B Pereira
A nossa bandeira
"Auri verde pendão de minha terra
Que a brisa do Brazil beija e balana
Estandarte que a luz do sol encerra
As promessas divinas da Esperança."
A bandeira brazileira a mais bonita de todas; vou descrevel-a. O rectangulo verde indica a cor de nossas mattas. O losango amarello indica a cor das riquezas naturais que o nosso caro Brazil encerra como o ouro. No centro da bandeira vł-se uma esphera azul que indica a terra, . . . Salve! Bandeira Brazileira

Statement

O documento foi retirado de uma exposição on-line de manuscritos do estado de São Paulo do início do século XX. Quanto à relevância social para o leitor da atualidade, o texto

Alternatives

(a) funciona como veículo de transmissão de valores patrióticos próprios do período em que foi escrito. [correct] (b) cumpre uma função instrucional de ensinar regras de comportamento em eventos cívicos. (c) deixa subentendida a ideia de que o brasileiro preserva as riquezas naturais do país. (d) argumenta em favor da construçção de uma nação com igualdade de direitos. (e) apresenta uma metodologia de ensino restrita a uma determinada época.

Figure 1. Example of original question (top) and machine-readable format (bottom). Some text was suppressed from the header to save space.

containing the text given as base knowledge for the question; the *statement*, containing the question's statement; and the *alternatives*, containing each answer candidate's text and flagging the correct one. Figure 1 depicts an example of a digital question and its machine-readable format in the structured form we propose. The picture contains a text written by a student describing the Brazilian flag, sketched on the upper left corner. Textual information that is irrelevant or not easily recognized as text is (at this stage) ignored from the conversion. In the example, this includes the reading of the stamp mark and the annotations in the picture.

To help in the analysis of the performance of techniques, we associate informative tags to questions. The tag "image" (IMG) is associated to every question that is accompanied by an image, regardless of whether it is actually important or crucial to answer the question. By image we considered anything that is not purely textual: drawings, pictures, graphics, tables, mathematical equations and diagrams. The remaining tags inform what kind of knowledge (or tasks) are (in principle) necessary to answer the question.

The tag "encyclopedic knowledge" (EK) suggests that the question resembles (or is) a factoid question, thus it can be answered by consulting an external source of knowledge such as an encyclopedia. This is in contrast with questions that can be answered only using the text or image (and commonsense knowledge and reasoning). Examples of questions tagged as EK include the characteristics of a social movement and the the main ideas of a philosopher.

The tag "image compreheension" (IC) is assigned to questions which require identifying or understanding the constituent elements of a given painting, cartoon, photo or advertisement. We remark that we included in this class images that contain text, even when it is only this text that might be crucial to producing the correct answer. An example is a question that displays a cartoon and then asks: "The cartoon criticizes the means of communication, specially the Internet, because", this one demands understanding what is inside this cartoon, therefore the answer lies in the image.

Note that if no graphical feature (font, text layout, etc.) is relevant to produce the answer then the question is tagged as TC and not as IC. An example is Figure 1, where the graphical features play no role in both statement or answers.

Finally, a question is tagged as "text comprehension" (TC) if the answer can be identified somehow using the given text. As the answer is seldom stated ipsis litteris, this tag tend to require some sort of reasoning about what is stated and frequently asks for identifying: (1) the author's thoughts or feelings; (2) figures of speech; (3) passages with some characteristics. These are usually highlighted by the presence of expressions such as "as the author", "present at the text fragment", etc. We present in Figure 2 an example of question with this format (our translation).

We point out that questions tagged as TC are the ones that require the most commonsense reasoning and understanding of Portuguese. Also, in the unused exams, viz. Sciences and Mathematics, it would be useful to specialize this knowledge into two others: (1) one stating the necessity to convert the given problem in natural language into a mathematical or chemistry formula, solve it and identify the most similar answer candidate; (2) one identifying that the question requires understanding domain specific rules in the given text, for domain specific rules we understand the Laws of Physics, Thermodynamics, and so on.

These tags are not mutually exclusive. For example, consider the question in Figure 3. The correct answer requires both text interpretation, context understanding and knowing basic facts. Questions tagged as IC and EK usually present an image of an event or person mentioned in the statement; questions on cartoons are usually TC and IC, since the text appears inside an image and the answer is in textual form.

Table 1 shows the overall number of questions in the exams of Humanities (1) and Languages (2) (discarding Foreign Language) from 2010 to 2015, as well as the number of questions associated with each tag. We see that (i)

Header

TEXT I

Our fight is for the democratization of land property, which is getting more and more concentrated in our country. Around 1% of all landowners controls 46% of the land. We pressure through occupations of big or unproductive land properties, that don't do their social part, as the Constitution of 1988 demands. We also occupy farms whose land was stolen from public land.

TEXT II

The small landowner is equal to a small store owner: the smaller the business, harder it is to keep it running, because the charges are heavy and it must profit. I am in favor of productive and sustainable properties that generate jobs. Supporting a productive, job generating enterprise is cheaper and generates much more than supporting land reform.

Statement

In each fragment the authors oppose each other. This happens because the authors associate the land reform, respectively, to

Alternatives

(a) reduction of city swelling and criticism on small land owners. (b) growth of national funds and prioritize the international market. (c) stopping the mechanization of agriculture and fighting the rural exodus. (d) privatization of state companies and economical growth stimuli. (e) correcting historical distortions and loss of agribusiness. [correct]

Figure 2. Example of question tagged as Text Comprehension. References were suppressed on the example, our translation

Header

Six p.m., Preciados Street. Far away, the human mass that fills the Puerta Del Sol Square in Madrid stands up. A group of girls, seeing this, runs towards the crowd. Millions of people shout the slogan: "Do not, do not, do not represent us". A boy speaks in the megaphone: "We demand a referendum about the bailout".

Statement

In 2011, the Spanish Indignados' encampment expressed the discontent of the European youth with the politicians. Which proposal synthesizes the set of political claims made by these young people?

Alternatives

(a)Universal Suffrage. (b)Direct Democracy.[correct] (c)Additional parties. (d)Autonomous legislation. (e)Parliamentary immunity.

Figure 3. Question requiring text comprehension (TC) and encyclopedic knowledge (EK), our translation

most questions require text comprehension, (ii) about 40% can be answered by consulting an external knowledge base, (iii) many questions that use irrelevant images. Note that in the database there is no tag of exclusivity, we are displaying here just to compare how many "pure" questions there are.

4. Baseline Methods

In this section, we investigate two approaches for solving the ENEM test, one based on information retrieval (IR) and other based on Word2Vec (W2V). The purpose of these methods is to serve as a baseline for the future, and to attest that solving the ENEM is not a trivial task, meaning that it requires some level of knowledge representation and reasoning.

The first strategy that one may think when trying to solve a question is to look for the words of the question and the answers in a text, this is what IR is about.

TABLE 1. USAGE OF EACH TYPE OF KNOWLEDGE ON HUMANITIES(1) AND LANGUAGES(2) FROM 2010 TO 2015

EXAM	#TOTAL	IMG	TC	EK	IC	TC$_{only}$	EK$_{only}$
2010-1	45	5	31	29	1	16	13
2010-2	40	9	31	9	8	25	3
2011-1	45	9	30	32	6	12	11
2011-2	40	12	29	11	11	21	2
2012-1	45	9	31	20	6	21	9
2012-2	40	13	35	11	10	23	3
2013-1	45	10	32	20	9	19	5
2013-2	40	12	35	10	10	23	0
2014-1	45	12	32	28	10	13	7
2014-2	40	8	34	12	7	22	3
2015-1	45	10	36	18	7	22	4
2015-2	40	11	35	9	8	23	1
Total	510	120	391	209	93	240	61

This approach consists in building a database of text documents, which is then used to answer questions by searching common words in the question, answer and database. This is usually done by building an inverted index of the text corpora, then scoring and retrieving documents by their similarity to a query document, where similarity is taken as number of matched words (i.e., words that occur in both the query and the retrieved document).

To answer a question we create, for each alternative i, a query q_i containing the text in statement s *augmented* with the text in the alternative a_i. We use each query to retrieve the top scoring document, and we select the alternative whose retrieved document obtained the highest score — in case of "indecision" we assume that an alternative is selected at random.

The IR approach has as limitation the necessity of matching the words used in the query with the exact same words in the database, but that's something that seldom happens, so it would be interesting to look not only to the overlap of words, but also to the similarity of the semantics.

W2V is a neural network based method for learning vector representations of words such that semantically similar words are represented by near vectors [22]. The word vector representation is learned by maximizing the cross entropy between the word and its context (words that co-occur in a predetermined size window). Among other feats, W2V have been shown to perform well in analogy tasks such as "Man is to King like Woman is to ?". This can be achieved by finding the closest vector to $vector(\text{king}) - vector(\text{man}) + vector(\text{woman})$ (which is expected to be $vector(\text{queen})$), where sum and subtraction operations are standard pointwise operations and dissimilarity is usually measured by the cosine distance. To answer questions using W2V we sum the vectors of every word in *header* and *statement* creating the vector V_q and, analogously, other 5 vectors V_i. The selected answer corresponds to the closest vector V_i to V_q (using cosine distance).

TABLE 2. Exam, number of questions used, accuracy (in percentage) and ranking (inside brackets) of each approach.

EXAM	USED	IR-H	IR-E	IR-W	NDH-E	AH	W2V
2010-1	40	26.5 (6)	30.0 (2)	27.5 (3)	27.5 (3)	27.5 (3)	40.0 (1)
2010-2	31	26.4 (4)	29.0 (3)	16.1 (6)	32.2 (1)	29.3 (2)	22.5 (5)
2011-1	36	34.4 (2)	25.0 (4)	22.2 (5)	36.1 (1)	30.5 (3)	22.2 (5)
2011-2	28	27.1 (2)	21.4 (4)	28.5 (1)	25.0 (3)	21.4 (4)	21.4 (4)
2012-1	36	25.5 (3)	22.2 (4)	22.2 (4)	27.7 (2)	22.2 (4)	33.3 (1)
2012-2	27	25.9 (4)	29.6 (2)	37.0 (1)	25.9 (4)	29.6 (2)	25.9 (4)
2013-1	35	22.2 (4)	25.7 (1)	25.7 (1)	20.0 (5)	25.7 (1)	17.1 (6)
2013-2	28	27.8 (2)	21.4 (5)	28.5 (1)	25.0 (3)	25.0 (3)	17.8 (6)
2014-1	33	22.4 (4)	24.2 (2)	21.2 (5)	21.2 (5)	27.2 (1)	24.2 (2)
2014-2	32	28.7 (3)	28.1 (4)	25.0 (5)	31.2 (1)	31.2 (1)	25.0 (5)
2015-1	35	24.5 (2)	17.1 (4)	22.8 (3)	25.7 (1)	17.1 (4)	17.1 (4)
2015-2	29	27.5 (3)	34.4 (1)	13.7 (6)	27.5 (3)	31.0 (2)	27.5 (3)
Average		26.5 (3.2)	24.9 (3.0)	24.2 (3.4)	27.0 (2.6)	26.4 (2.5)	24.5 (3.8)
SD		3.1	4.7	6.1	4.4	4.4	6.7

5. Empirical Results

We used the Lucene software to efficiently index and retrieve documents from two different corpora: a corpus of 1262 documents of ten thousand lines extracted from the Wikipedia; and a corpus of questions containing ENEM exams between 2009 and 2015 (each document consists of the text of the header, statement and correct answer of a question). When using the ENEM corpus, we hold out the exam which we are solving from the database.

We trained the W2V model using the same Wikipedia corpus. We also experimented with a third corpus, which is created by augmenting the ENEM model as follows. When solving a question, for every word w in the query we add the most similar word to w according to W2V. We thus solved each exam using either the "regular" queries as described in the previous section (i.e., without the extra words added by W2V), and using the "augmented" queries.

For the IR approach, we experimented with three different strategies to build a corpus of text documents. The IR-H strategy uses only the header of the question to be solved to build the corpus of documents to be retrieved. For this corpus will have always only one document, this strategy often finds no documents matching the query, creating thus a case of complete indecision. We reward these indecision cases with 0.2 points (corresponding to the expected value of selecting an alternative at random). The IR-E strategy uses the entire text of the question to build the corpus: header, statement and correct alternative. Finally, the IR-W strategy uses articles from Wikipedia as corpus.

We also evaluated two different heuristics for fusing the results of IR heuristics. The **Adding Heuristic** (AH) adds the score of each alternative according to IR-E and IR-H, and then selects the top scoring alternative. We have tried combining other strategies (e.g., IR-E and IR-W) but they did not produce good results and are omitted for the sake of space. The **Non-Deciding Heuristic** (NDH) uses the same answer given by IR-H if the top scoring answer is unique, otherwise it selects the solution of some other strategy. We tried with IR-E and IR-W for breaking ties, obtaining heuristics NDH-E and NDH-W, respectively. We have also tried combining the results of AH and NDH but this did not lead to better results (and are omitted).

Table 2 shows the percentage of correct answers and ranking of all methods using only questions with no image (the "used" column indicates the number of questions per exam), for each exam we counted the number of hits using normal queries and using the augmented queries, from these two we select the highest and present its percentage. NDH-E and NDH-W performed very similarly, so we report only the former. We see that all approaches outperformed random guessing on average ($> 20\%$), and that the accuracy standard deviation was moderate. In terms of average accuracy, NDH scored higher, followed by IR-H and AH. AH had the lowest mean rank (i.e., the best), followed by NDH and IR-E, showing the effectiveness of combining different information sources. By inspecting the table, we see that only IR-H and NDH were consistently better than random guessing in all exams. IR-H accuracy was very consistent (with the lowest standard deviation). This suggests that headers have enough information to solve questions, from which text understanding techniques can greatly benefit.

TABLE 3. Best scores in percentage of each heuristic when solving each type of knowledge

TAG	IR-H	IR-E	IR-W	NDH-E	NDH-W	AH	W2V
EK	28.5	26.7	27.2	31.5	29.1	28.2	29.6
EK$_{only}$	30.4	36.0	37.7	31.1	34.0	31.1	22.9
TC	25.4	24.0	24.8	25.5	26.5	26.0	25.3
TC$_{only}$	26.9	24.5	24.5	26.6	27.0	28.3	22.9
IC	22.5	18.2	27.9	20.4	26.8	20.4	26.8

We also note that even though IR-W used the by far largest corpus, it was outperformed by all other methods. This may be due to the fact that information in Wikipedia are too broad and not strictly pertinent to topics of exam, which might actually hurt the performance of IR methods. To test that, we ran some preliminary tests using only part of the Wikipedia corpus (we tested with several partitions); the results did not indicate a significant improvement.

The W2V approach, while competitive in terms of average accuracy, was very inconsistent across exams, obtaining a high standard deviation and a high mean rank. It achieved the highest accuracy in a exam among all approaches (viz. 40% in 2010-1), but it often performed worse or only slightly better than random guessing (2011-2, 2013, 2015-1). In Table 3, we segment the accuracies of each approach by the type of knowledge required to solve questions. The numbers report the normalized percentage of hits (no. of correct answers divided by total of questions of that type).

IR-W outperforms IR-E in questions requiring external knowledge but no text comprehension, while both perform similarly when question require only text comprehension. The larger corpus used by IR-W seems to give an advantage when solving questions that require image comprehension. This is of course not due to any image processing (which we did not perform), but likely because the accompanying text contains general clues about the image itself, which can be better exploited using a larger and broader corpus.

The two variants of NDH differ significantly for some types of knowledge, and were thus distinguished. We see that NDH-E performed slightly better than random in IC questions, while NDH-W was able to sustain its aboverandom-guessing performance across different tags. This is a consequence of the superiority of IR-W over IR-E on such questions. We also see that in questions requiring only EK both NDH-E and NDH-W perform worse than IR-E and IR-W, respectively. AH also did not improve in questions requiring only EK, but it did better than individual heuristics in questions requiring only TC. The W2V was not competitive in questions requiring only either EK or TC, but performed similarly to NDH-W in IC questions. Finally, we see that all approaches, except W2V, performed better in questions that require only EK, which suggests that these questions might require less reasoning skills than others.

6. Conclusion

In this work, we propose using the Exame Nacional do Ensino Médio (ENEM), a High School level exam widely used by Brazilian Universities as entrance exam, as a guiding test for AI. We described how a machine-readable database of questions from previous editions of ENEM is being built, and presented and evaluated some baseline approaches based on information retrieval and word vector representation techniques. Among other qualities, the proposed test consists of a real-word task with genuine interest from the public, uses a limited but very broad domain, has a fair and clear evaluation scheme by means of percentage of correct answers, allows comparison against human performance, and requires commonsense knowledge and reasoning. Our hope is that this test will attract interest from the wide public as well as from AI researchers, and foster the development of interesting solutions in natural language processing that can be also used outside Portuguese-oriented tasks, image processing and knowledge representation and reasoning.

Although data about the average score obtained by students who took the exam are available, they are not directly comparable to our results, since questions are weighted differently, and the corresponding weights remain secret. Therefore, a direct comparison of automated methods and human performance is not yet possible.

In the future we plan to compare results with human performance (by asking volunteers to take the test), incorporate all questions in the database, and develop more sophisticated techniques that perform some sort of text understanding, probably borrowing insight from the field of question answering.

Acknowledgments

This work was partly supported by CAPES, CNPq (grants #303920/2016-5, #420669/2016-7) and FAPESP (grant 2016/01055-1).

References

[1] S. M. Shieber, "Does the Turing test demonstrate intelligence or not?" in *Proc. 21st National Conf. Artif. Intell. and the 18th Innovative App. Artif. Intell. Conf.*, 2006, pp. 1539–1542.

[2] J. R. Searle, "Minds, brains, and programs," *Behav. and Brain Sci.*, vol. 3, pp. 417–424, 1980.

[3] N. Block, "Psychologism and behaviorism," *Philo. Rev.*, vol. 90, pp. 5–43, 1981.

[4] H. J. Levesque, "The winograd schema challenge," in *Logical Formalizations of Commonsense Reasoning, 2011 AAAI Spring Symp.*, 2011.

[5] S. M. Shieber, "Lessons from a restricted turing test," *Commun. ACM*, vol. 37, no. 6, pp. 70–78, 1994.

[6] K. Warwick and H. Shah, *Turing2014: Tests at The Royal Society, June 2014*. Cambridge University Press, 2016, pp. 171–186.

[7] "Turing test success marks milestone in computing history," (http://www.reading.ac.uk/news-and-events/releases/PR583836.aspx), 2015, [Online; accessed April-23-2017].

[8] H. Kitano, M. Asada, Y. Kuniyoshi, I. Noda, and E. Osawa, "Robocup: The robot world cup initiative," in *Proc. 1st Int. Conf. on Autonomous Agents*, 1997, pp. 340–347.

[9] R. Gerndt, D. Seifert, J. H. Baltes, S. Sadeghnejad, and S. Behnke, "Humanoid robots in soccer: Robots versus humans in robocup 2050," *IEEE Robot. Automat. Mag.*, vol. 22, no. 3, pp. 147–154, 2015.

[10] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty, "Building watson: An overview of the DeepQA project," *Artif. Intell. Magazine*, vol. 31, no. 3, 2010.

[11] E. Davis, "How to write science questions that are easy for people and hard for computers," *Artif. Intell. Magazine*, vol. 37, no. 1, pp. 13–22, 2016.

[12] Y. Miyao and A. Kawazoe, "University entrance examinations as a benchmark resource for NLP-based problem solving," in *Proc. Int. Joint Conf. on Natural Language Processing*, 2013, pp. 1357–1365.

[13] P. Clark, P. Harrison, and N. Balasubramanian, "A study of the knowledge base requirements for passing an elementary science test," in *Proc. 2013 Workshop on Automated Knowledge Base Construction*, ser. AKBC '13, 2013, pp. 37–42.

[14] T. Khot, N. Balasubramanianm, E. Gribkoff, A. Sabharwal, P. Clark, and O. Etzioni, "Markov logic networks for natural language question answering," *Proc. Conf. on Empirical Methods in Natural Language Processing*, p. 685694, 2015.

[15] P. Clark, O. Etzioni, T. Khot, A. Sabharwal, O. Tafjord, P. D. Turney, and D. Khashabi, "Combining retrieval, statistics, and inference to answer elementary science questions." AAAI Press, 2016, pp. 2580–2586.

[16] S. Matsuyoshi, Y. Miyao, T. Shibata, C.-J. Lin, C.-W. Shih, Y. Watanabe, and T. Mitamura, "Overview of the ntcir-11 recognizing inference in text and validation (rite-val) task." Proc. 11th NTCIR Conf., 2014, pp. 223–232.

[17] A. Fujita, A. Kameda, A. Kawazoe, and Y. Miyao, "Overview of todai robot project and evaluation framework of its nlp-based problem solving," in *Proc. 9th Int. Conf. on Language Resources and Evaluation*, 2014.

[18] D. Santos, "Porquł o Pgico? Razes para uma avaliao conjunta," *Linguamtica*, vol. 4, no. 1, pp. 1–8, 3 de Abril 2012.

[19] N. Cardoso, "Medindo o precipcio semntico," *Linguamtica*, vol. 4, no. 1, pp. 41–48, 3 de Abril 2012.

[20] R. Rodrigues, H. G. Oliveira, and P. Gomes, "Uma abordagem ao págico baseada no processamento e análise de sintagmas dos tópicos," *Linguamtica*, vol. 4, no. 1, pp. 41–48, 3 de Abril 2012.

[21] "Provas e gabaritos," Instituto Nacional de Estudos e Pesquisa, http://portal.inep.gov.br/provas-e-gabaritos, [Online; acessed April-23-2017].

[22] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, ser. NIPS'13, 2013, pp. 3111–3119.