Distractor quality evaluation in Multiple Choice Questions

Van-Minh Pho, Anne-Laure Ligozat, Brigitte Grau

Distractor quality evaluation in Multiple Choice Questions

Van-Minh Pho[1,2], Anne-Laure Ligozat[1,3], and Brigitte Grau[1,3]

[1] LIMSI-CNRS, Orsay, France
[2] Université Paris-Sud, Orsay, France
[3] ENSIIE, Evry, France
prenom.nom@limsi.fr

Abstract. Multiple choice questions represent a widely used evaluation mode; yet writing items that properly evaluate student learning is a complex task. Guidelines were developed for manual item creation, but automatic item quality evaluation would constitute a helpful tool for teachers.
In this paper, we present a method for evaluating distractor (i.e. incorrect option) quality that combines syntactic and semantic homogeneity criteria, based on Natural Language Processing methods. We perform an evaluation of this method on a large MCQ corpus and show that the combination of several measures enables us to validate distractors.

Keywords: Multiple choice questions, natural language processing, automatic quality evaluation, production of educational material

1 Introduction

Multiple Choice Questions (MCQs) are widely used in many educational and evaluation contexts since their assessment can be automated and they have proven to be relevant and objective indicators of the learner skills [5]. Yet, writing multiple choice items is costly, and the quality of MCQ is crucial in order to ensure that learners will perform according to their skills. Guidelines have been developed [1] to help create quality MCQs. However, actual MCQs can present flaws because educators may not have formal instruction for writing MCQs [16]. An automatic evaluation of MCQs quality could thus help educators.

A MCQ is composed of two parts (cf. example below): the *stem* and the *options* (or *choices*), which include both the answer (correct option), and one or several *distractors* (incorrect options).

Stem: What country is Kimchi from?
Answer: Korea
Distractors: Japan, China, Mongolia

Selecting distractors is a difficult task when creating a MCQ: the quality of a MCQ relies heavily on the quality of these options [15].

The objective of this paper is the automatic evaluation of the quality of options according to writing rules. In particular, the rule *"Keep choices homogeneous in content and grammatical structure"* leads us to propose a definition of syntactic and semantic homogeneity. Homogeneous in content means that options must share semantic features, but are, nonetheless, sufficiently different to be plausible answers but not possible answers.

We propose to evaluate homogeneity by comparing each distractor to the answer and by computing criteria based on syntactic parsing and the use of different semantic resources to cover numerous semantic relationships that are combined in a machine learning model. We focused our work on short options, expressed by chunks (lowest-level phrases) and named entities. Previous research on automatic distractor selection had more restrictions on the type of options [6, 11], or was dedicated to a specific domain [6]. To our knowledge, it is the first attempt to evaluate automatically the semantic quality of MCQ distractors. We will show that our method outperforms state-of-the-art methods.

2 Evaluating the homogeneity of options

In order to help teachers create well-formed MCQs, guidelines were developed. One of the most popular set of guidelines was written by [5]. It is composed of 43 rules grouped in categories related to MCQ content, MCQ formatting, MCQ style, stem writing and option writing.

Concerning option writing, the most important ones for quality evaluation are the following: *"Keep choices homogeneous in content and grammatical structure"* and *"Keep choices independent; choices should not be overlapping"*.

Grammatical homogeneity can be verified on the syntactic representation of the options provided by a natural language (NL) parser. In the example of Section 1, all options are noun chunks. Semantic characterization of options can be based on different NL tools, each able to provide different semantic properties according to the resources they are based on.

To evaluate the quality of distractors, we compare them to the answer, both on syntactic and semantic features. We do not want to learn a decision that would state on distractor homogeneity or not because homogeneity is more a question of degree than a binary decision. Thus, we formulate homogeneity evaluation as a ranking problem: the most homogeneous distractors should be classified in the first ranks. The candidates to rank, except for distractors, are selected according to syntactic homogeneity criteria, which proved to be a valid selection criteria in the corpus study of [12].

3 Semantic homogeneity

Semantic homogeneity states that options share common semantic characteristics. In the example of Section 1, all options are Asian countries.

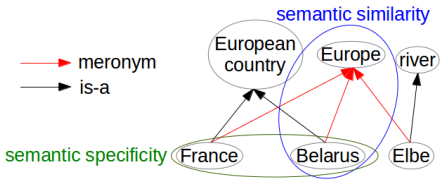

Fig. 1. Semantic characterization of pairs of nodes

We will define several notions close to semantic homogeneity by considering a knowledge organization in the form of a hierarchical graph (Figure 1) carrying typed concepts referred by the terms of the options and semantic relations.

The definition of *semantic relatedness* by [13] is the following: "Semantic relatedness indicates how much two concepts are semantically distant in a network or taxonomy by using all relations between them (i.e. hyponymic/hypernymic [4], antonymic [5], meronymic [6] and any kind of functional relations including is-made-of, is-an-attribute-of, etc.)". Thus, semantic relatedness holds between two concepts when there exists a path between them, and the degree of relatedness is dependent on the path length and the types of relations. In Figure 1, all concepts are semantically related.

We define *semantic similarity* as a particular case of semantic relatedness: two terms are similar if they share the same meaning (i.e. synonyms) or a partial meaning, i.e. the concepts to which they refer are linked by a purely ascending or descending chain of is-a or meronymy relations, as "Belarus" and "Europe" in Figure 1.

We define *semantic specificity* as a particular case of semantic relatedness between two concepts that share a common ancestor, as "France" and "Belarus" in Figure 1.

We define *semantic homogeneity* as a particular case of semantic relatedness: it considers all relations between compared concepts. Moreover, semantic homogeneity excludes the notion of semantic similarity: two options cannot be similar. Finally, a better homogeneity is reached if semantic specificity is respected.

To estimate semantic homogeneity between a distractor and the answer, we compute several semantic relatedness scores. These measures are based on different criteria: measures based on hierarchical semantic representations of concepts, and data driven measures based on contextual relatedness, i.e. the principle stating that terms with similar contexts in a corpus are likely to be semantically related. Hierarchical representations allow us to take into account explicit semantic relations to compare two concepts. However, the resources (DBpedia, WordNet) corresponding to these representations are often limited in their cov-

[4] Two concepts of which the first has a more specific/general sense than the second
[5] Two concepts of which their senses are opposite
[6] Two concepts of which the first is a part or a member of the second

erage (proportion of options present in the resources). Measures based on contextual relatedness have a broader coverage, but the nature of semantic relations is unknown.

In the following sections, we present each of these measures.

3.1 Identity of named entity types

A named entity (NE) is an particular expression referred by a semantic class called NE type. Two options annotated with the same NE type are semantically specific, if they do not refer to a same concept. In order to measure their specificity, we consider 3 large categories: *location*, *organization* and *person*. For such types, NE recognizer are based on surface criteria and gazetteers and do not require a semantic knowledge base.

To compare the NE type of two terms, we use the following measure:

$$same\_NE\_type(t_1, t_2) = \begin{cases} 1 \text{ if } NE(t_1) = NE(t_2) \wedge t_1 \text{ is a NE} \wedge t_2 \text{ is a NE} \\ 0 \text{ else} \end{cases} \tag{1}$$

where t_1 and t_2 are two terms and $NE(t)$ is the NE type of the term t.

3.2 Similarity of semantic types provided by DBpedia

In addition to comparing general NE types, we compare semantic types at a more fine-grained and hierarchical level, which allow us to verify more precisely semantic specificity. However, while NE types can be recognized independently of a resource, specific types have to be recognized for concepts belonging to a hierarchical taxonomy. We chose DBpedia [7], a hierarchical resource built from Wikipedia articles. DBpedia entities are associated with semantic types which represent classes of the DBpedia ontology, organized in a taxonomy [8].

To compute semantic homogeneity between two terms t_1 and t_2 based on their DBpedia type and position in the taxonomy, we use Wu and Palmer's measure [17], $wup(t_1, t_2)$, which is based on is based on the shortest path between two concepts weighted by their depth in the taxonomy.

$$wup(t_1, t_2) = \frac{2 \times depth(lcs)}{depth(type(t_1)) + depth(type(t_2))} \tag{2}$$

where $type(t)$ is the DBpedia type of the term t, $depth(u)$ is the depth of a type u in the taxonomy and $lcs(type(t_1), type(t_2))$ is the least common subsumer (in terms of path length in the taxonomy) between $type(t_1)$ and $type(t_2)$. Thus, two deep concepts with a common parent get a higher score than two less deep concepts with a common parent.

[7] http://dbpedia.org/About
[8] http://mappings.dbpedia.org/server/ontology/classes/

3.3 Relatedness measures based on WordNet

To measure semantic homogeneity for all kinds of options and in particular non NE ones, we use measures defined on WordNet [9], a lexical network that clusters synonym words in *synsets* linked by semantic relations. To each synset is associated a gloss, i.e. a definition in NL. WordNet also contains named entities for a few kinds of entities (large cities, countries, continents...). We use the four measures selected by [11]: the extended gloss overlap measure based on textual similarity between the glosses of two concepts; Leacock and Chodorow's measure based on the shortest path between concepts; and Jiang and Conrath's and Lin's measures, both based on *information content* [14]. Terms can have multiple senses, so they can be associated with multiple synsets. To compute semantic relatedness between two terms, we compute the measures on all pairs of synsets associated with the terms and we keep the maximal score.

3.4 Comparison of links of Wikipedia articles

We also considered measures based on contextual relatedness. A possible contextual representation of a term is the sets of incoming and outgoing links associated with a page in Wikipedia. We consider pages whose title corresponds to an option. The incoming and outgoing "manually created" links represent associated concepts. The tool Wikipedia Miner [10] computes a score learned on these links from Wikipedia dumps.

3.5 Explicit Semantic Analysis

Another contextual representation of terms is their distribution through documents in a corpus. Two terms having close distributions in the same documents are likely to be semantically related. In order to compare the distributions of a candidate and the answer, we computed a measure based on Explicit Semantic Analysis (ESA) [4]. ESA is based on a vector representation of texts in which the dimensions are the weights of the text in each document belonging to a corpus. A word is represented by a vector containing weights on its frequency in each document and a text is represented by the centroid of the weighted vectors representing each word of the text. The relatedness score of two texts is the cosine of the vectors representing these texts. In our case, the document corpus is Wikipedia. To compute the measure based on ESA, we use the tool ESAlib [10].

4 Evaluation of distractor quality

In order to evaluate the quality of distractors, we merge existing distractors with non-distractors (terms which have not been manually selected to be distractors

[9] http://wordnet.princeton.edu/
[10] http://ticcky.github.io/esalib/

for the MCQ). Our purpose is to learn an assessment model able to rank distractors above non-distractors, as they should be more homogeneous to the answer than the non-distractors. None of the proposed measures directly estimates semantic similarity but we made the hypothesis that a high score of semantic relatedness can represent semantic similarity, which should be learned by our model. MCQs that we process are associated with a document from which stems have been conceived. Non-distractors are selected in this document according to syntactic homogeneity. A first step involves annotation of the options and the non-distractors.

4.1 Document and options annotation

To extract non-distractors and compute the different measures, candidates and answers have to be annotated by syntactic and semantic information, which is better realized if these text excerpts are analyzed in the reference document. Thus, we perform four annotations of the document, in the following order:

1. syntactic parsing with the Stanford Parser [7];
2. NE annotation with the Stanford Named Entity Recognition tool [3];
3. specific type annotation, to find entities which are related to a DBpedia entity (and, by extension, a Wikipedia article), with DBpedia Spotlight [2]. This tool associates DBpedia entities with corresponding entities of the document and disambiguates these entities if required. However, some terms (chunks and/or NEs) are not annotated by DBpedia Spotlight. We associate these terms with all DBpedia entities whose title corresponds to them, so without disambiguation;
4. WordNet concept annotation, aiming at associating terms with a WordNet concept. This annotation is performed on chunks and/or NEs, i.e. single or multiword expressions, as following:
   - if the expression appears in WordNet, the expression is associated with its corresponding concept;
   - if the expression does not appear in WordNet and is not a NE, the expression is associated with the concept corresponding to its syntactic head (for instance, the expression "the little cat" is associated with the WordNet concept "cat").

Annotations of the options are extracted from their occurrences in the document. If an option does not appear in the document, its annotations are performed similarly to the document.

4.2 Non-distractor extraction and annotation

Since MCQs are related to a reference document, non-distractors are extracted from this document. All non-distractors are syntactically homogeneous to the answer. If the answer is a NE, the non-distractors are the NEs of the reference document, as [12] showed that distractors generally have the same NE type as

corpus	set	# q.	# opt.	# opt./q.	% allMCQ	purpose
mcqNE	qa4mre	56	252	4.5		machine reading
	englishEval	47	150	3.2		language evaluation
	total	103	402	3.9	14	
mcqNonNE	qa4mre	51	239	4.7		machine reading
	englishEval	100	342	3.8		language evaluation
	total	151	581	3.8	20	

Table 1. Characteristics of the corpora

the answer. Nevertheless, in order to take metonymy into account, we keep all non-distractors with a NE type, regardless of the type. If the answer is a chunk and not a NE, the non-distractors are the chunks of the reference document with the same syntactic type as the answer. The chunks are selected from the parse tree of the document sentences, with Tregex [9], a tool that selects nodes in parse trees from patterns. We associate non-distractors with their NE type, DBpedia entity and WordNet concept annotated in the document. A last filtering consists of removing non-distractors similar to an option, in order to avoid overlaps: two elements are considered similar if they are associated with the same DBpedia entity or if they refer to the same synset in WordNet.

4.3 Semantic ranking

Ranking of the candidates (distractors and non-distractors) according to the different criteria of semantic homogeneity is performed by the tool SVMRank [11], an automatic ranker based on a SVM model, that compares couples of distractors-non-distractors of a same MCQ and learns the weights of the criteria such as for each couple of distractor-non-distractor (d, nd), $svm(d) > svm(nd)$.

5 Experiments

5.1 Data sets

In order to evaluate our method, we use an English corpus of MCQs (corpus allMCQ, 735 MCQs) extracted from different sources: machine reading system tests provided by QA4MRE [12] (set qa4mre) and several websites of English language learning (set englishEval). We assume that these MCQs tend to be well-written. From this corpus, we established two sub-corpora: the first is composed of MCQs which answers are NEs (corpus mcqNE, 14 % of allMCQ), and the second is composed of MCQs which answers are chunks and not NE (corpus mcqNonNE, 20 % of allMCQ).

The questions that we process (chunks and NEs) compose more than one third of the original corpus which shows that these types of questions are frequently asked in tests. Learning was performed separately on each sub-corpus.

[11] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html
[12] http://www.celct.it/newsReader.php?id_news=74

	mcqNE			mcqNonNE		
	R	P	F	R	P	F
NE similarity	0.83	0.26	0.40			
DBpedia type similarity	0.70	0.34	0.46	0.94	0.14	0.24
Extended gloss overlap	0.67	0.25	0.36	0.37	0.23	0.28
Leacock & Chodorow	0.73	0.27	0.39	0.42	0.22	0.29
Jiang & Conrath	0.83	0.23	0.36	0.40	0.18	0.25
Lin	0.84	0.23	0.36	0.40	0.18	0.25
Wikipedia link similarity	0.41	0.32	0.36	0.76	0.22	0.34
ESA	0.40	0.34	0.37	0.35	0.24	0.28
Combination	0.48	**0.46**	**0.47**	0.39	**0.36**	**0.37**

Table 2. Results of semantic relatedness methods

5.2 Evaluation methodology

We consider that distractors are semantically closer to the answer than non-distractors and should thus have a higher rank. In order to evaluate candidate ranking, we compute the average precision (Equation (3)) and recall (Equation (4)), as well as the f-measure (Equation (5)).

$$AP = \frac{\sum_i^{nbQ} P_{i,nbD}}{nbQ} \quad (3) \qquad AR = \frac{\sum_i^{nbQ} R_{i,nbD}}{nbQ} \quad (4)$$

$$F = 2 \times \frac{AP \times AR}{AP + AR} \quad (5)$$

where nbQ is the number of MCQs in the corpus, nbD the number of distractors of the evaluated MCQ, and $P_{i,nbD}$ and $R_{i,nbD}$ are the precision (Equation (6)) and recall (Equation (7)) of MCQ i.

$$P_{i,nbD} = \frac{\#\text{D of rank} \leq nbD}{\#\text{C of rank} \leq nbD} \quad (6) \qquad R_{i,nbD} = \frac{\#\text{D of rank} \leq nbD}{nbD} \quad (7)$$

where D means distractors and C means candidates.

Precision and recall are computed for each semantic relatedness measure, as well as for the ranking model. We evaluate the ranking by 7-fold cross-validation.

5.3 Results

Table 2 shows that the ranking model gives higher balance between recall and precision than individual measures, regardless of the corpus. In particular it gives a higher precision than other measures and better results than WordNet-based measures, used by [11].

Some measures give a higher recall than the ranking model. We distinguish two cases: the first concerns (NE and specific) type-based measures which are more efficient for filtering candidates than selecting distractors. The second case concerns measures whose resource coverage is low (WordNet in mcqNE and Wikipedia in mcqNonNE).

The results are overall lower in the mcqNonNE corpus. The main reason is that non NE candidates and answers are associated with less semantic information than NE, particularly on semantic types.

In the corpus mcqNE, most cases where non-distractors are better ranked than distractors are due to the fact that distractors and answers are not typed by a very specific (DBpedia) type. The remaining non-distractors are relevant enough to be distractors or are similar to the answer, so cannot be distractors.

The majority of non-distractors of the corpus mcqNonNE which are better ranked than distractors are clearly non-distractors but some measures (particularly WordNet-based) consider that these non-distractors are more semantically related than distractors. Among the remaining non-distractors, some of them are not semantically related to the answers in the current context (reference document) or they are relevant enough to replace distractors.

6 Related work

Automatic distractor selection is usually based on similarity measures between the candidates and the answer and is evaluated by learners or teachers.

Existing work on automatic distractor selection is based either on hierarchical domain-specific resources (WordNet, UMLS) [6, 11] and/or document corpora [8, 11]. From these resources, candidates are selected according to common syntactic and semantic characteristics with the answer: common syntactic type [6, 8, 11], common semantic classes [6, 11] or terms sharing the same head as the answer [11]. Then, distractors are selected from candidates according to different measures: context-based [6, 8] or a strategy based on these first measures, WordNet-based and phonetic-based measures [11]. Evaluation of distractors is performed by learners (through psychometric tests) [8, 11] or judgment of domain experts [6], but none of this work evaluated distractors on a reference corpus. Moreover, related work is dedicated to a specific domain (linguistics, medicine, preposition learning), whereas our work is not specific to a domain. Related work is also limited by the syntactic types of answers (words, noun chunks) whereas our work covers all kinds of chunks (noun and verb phrases) and NEs.

7 Conclusion

In this paper, we proposed a method to automatically evaluate the quality of distractors according to criteria relative to syntactic and semantic homogeneity. Results outperform the state-of-the-art methods for automatic distractor selection, and are better on NE than other kinds of chunks. Measures based on hierarchical semantic resources allow us to filter candidates according to properties like types and semantic relations. Measures based on contextual relatedness allow us to refine distractor recognition.

These criteria are relevant but are not sufficient to automatically recognize distractors: considering other information like stems and the context of the options, we would recognize distractors more precisely. Moreover, in future work,

we will adapt our approach to all kinds of answers (phrases, clauses and sentences). We will also evaluate distractor quality *a posteriori*, from scores obtained by learners answering to MCQs.

References

1. Burton, S.J., Sudweeks, R.R., Merrill, P.F., Wood, B.: How to prepare better multiple-choice test items: Guidelines for university faculty. (1991)
2. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. Proceedings of SEMANTICS, 121–124 (2013)
3. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. Proceedings of the 43rd Annual Meeting on ACL, 363–370 (2005)
4. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. IJCAI, 7, 1606–1611 (2007)
5. Haladyna, T.M., Downing, S.M., Rodriguez, M.C.: A review of multiple-choice item-writing guidelines for classroom assessment. Applied measurement in education, 15 (3), 309–333 (2002)
6. Karamanis, N., Ha, L.A., Mitkov, R.: Generating multiple-choice test items from medical text: A pilot study. Proceedings of NLG, 111–113 (2006)
7. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. Proceedings of ACL, 1, 423–430 (2003)
8. Lee, J., Seneff, S.: Automatic generation of cloze items for prepositions. INTERSPEECH, 2173–2176 (2007)
9. Levy, R., Andrew, G.: Tregex and Tsurgeon: tools for querying and manipulating tree data structures. Proceedings of LREC, 2231–2234 (2006)
10. Milne, D., Witten, I.H.: An open-source toolkit for mining Wikipedia. Artificial Intelligence, 194, 222–239 (2013)
11. Mitkov, R., Ha, L.A., Varga, A., Rello, L.: Semantic similarity of distractors in multiple-choice tests: extrinsic evaluation. Proceedings of the EACL Workshop GEMS, 49–56 (2009)
12. Pho, V.-M., André, T., Ligozat, A.-L., Grau, B., Illouz, G., François, T.: Multiple Choice Question Corpus Analysis for Distractor Characterization. Proceedings of LREC, 4284–4291 (2014)
13. Ponzetto, S.P., Strube, M.: Knowledge derived from wikipedia for computing semantic relatedness. JAIR, 30, 131–212 (2007)
14. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007 (1995)
15. Rodriguez, M.C.: Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educational Measurement: Issues and Practice, 24 (2), 3–13 (2005)
16. Tarrant, M., Ware, J.: Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Medical Education, 42 (2), 198–206 (2008)
17. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. Proceedings of ACL, 133–138 (1994)