Entropy as a Proxy for Gap Complexity in Open Cloze Tests

Mariano Felice **Paula Buttery**
ALTA Institute
Computer Laboratory
University of Cambridge
Cambridge, UK
{mf501,pjb48}@cam.ac.uk

Abstract

This paper presents a pilot study of entropy as a measure of gap complexity in open cloze tests aimed at learners of English. Entropy is used to quantify the information content in each gap, which can be used to estimate complexity. Our study shows that average gap entropy correlates positively with proficiency levels while individual gap entropy can capture contextual complexity. To the best of our knowledge, this is the first unsupervised information-theoretical approach to evaluating the quality of cloze tests.

1 Introduction

Fill-in-the-gap or *cloze* test exercises are common means of assessing grammar and vocabulary in the realm of English as a Foreign Language (EFL). The most common example is the *multiple choice question*, which presents the student with a gapped sentence and a set of possible answers from which the right one is to be selected. These are referred to as *closed* cloze questions, since the answer is limited to the alternatives given. On the contrary, *open* cloze questions do not provide predefined options, so the student must produce an answer from scratch.

Generating these exercises is a laborious process, since they must be carefully designed to ensure they test the desired learning objective and do not confuse or present trivial questions to the student. For this reason, choosing the optimal locations in a sentence to insert the gaps and defining a suitable set of answer options becomes crucial, especially when exercises are generated automatically.

In this paper, we focus on open cloze tests and show how entropy can be used to assess the complexity of each gap in the text. Entropy is shown to provide insights into the expected difficulty of the question and correlate directly with the target proficiency level of the exercises. Exploiting this information should thus facilitate the automatic generation of more reliable open cloze exercises.

2 Related Work

Work on automated cloze test generation has mostly focused on multiple choice questions and distractor selection (Mitkov and Ha, 2003; Sumita et al., 2005; Brown et al., 2005; Lee and Seneff, 2007; Lin et al., 2007; Smith et al., 2010; Sakaguchi et al., 2013). Conversely, there has been little work on open cloze tests. Pino et al. (2008) describes a strategy to generate open cloze questions using example sentences from a learners' dictionary. Sentences are chosen based on four linguistic criteria: (grammatical) complexity, well-defined context (collocations), grammaticality and length. Further work improved on this method by providing hints for the gapped words (Pino and Eskenazi, 2009).

Malafeev (2014) developed an open source system to emulate open cloze tests in Cambridge English exams based on the most frequent gapped words. Expert EFL instructors found the generated gaps to be useful in most cases and had difficulty differentiating automated exercises from authentic exams. More recently, Marrese-Taylor et al. (2018) trained sequence labelling and classification models to decide where to insert gaps in open cloze exercises. The models achieved around 90% accuracy/F_1 when evaluated on manually created exercises.

While the quality of the generated gaps has traditionally been judged by human experts (Pino et al., 2008; Malafeev, 2014) or estimated from student responses (Sumita et al., 2005; Brown et al., 2005; Skory and Eskenazi, 2010; Beinborn et al., 2014; Susanti et al., 2016), systems should

Proceedings of Recent Advances in Natural Language Processing, pages 323–327,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_037

ideally predict the quality of the gaps during the generation process. In this regard, Skory and Eskenazi (2010) observe that Shannon's information theory (Shannon, 1948) could be used to estimate the reading difficulty of answers to a gap based on their probability of occurrence. Thus, for the sentence *"She drives a nice _____"*, the word "car" would be the most likely answer (lowest readability level) while words such as "taxi", "tank" and "ambulance" would be at increasingly higher levels.

Research on predicting the difficulty of cloze tests is also directly relevant to this work. Beinborn et al. (2014) built models to predict the difficulty of C-tests (i.e. gaps with half of the required word removed) at the gap and test level and later extended their approach to cover closed cloze tests (Beinborn et al., 2015; Beinborn, 2016). More recently, Pandarova et al. (2019) presented a difficulty prediction model for cued gap-fill exercises aimed at practising English verb tenses while Lee et al. (2019) investigated how difficulty predictions could be manipulated to adapt tests to a target proficiency level. Unlike our work, however, all these approaches are supervised and not applied to open cloze tests.

3 Entropy

In this paper, we build on the assumption that the complexity of a gap is correlated to the number of possible answers determined by the surrounding context and the likelihood of each answer. As noted by Pino et al. (2008), high-quality open cloze questions should sufficiently narrow the context of each gap in order to avoid multiple valid answers, which would make the exercise too broad in scope and therefore ineffective. We thus assume that gaps with more restricted context eliciting very specific answers should be more useful than broad gaps with very general answers, so the less "branching" that a gap allows, the better.

This property can be modelled by *entropy*, which quantifies the amount of information conveyed by an event. Intuitively, entropy can be considered a measure of disorder, uncertainty or surprise. If the probability of an event is very high, entropy will be low (i.e. there is less surprise about what will happen) while events with low probabilities will lead to higher entropy. Shannon's entropy, a common formulation to measure the number of bytes needed to encode information, is shown in Equation 1, where $P(x_i)$ stands for the probability of event x_i, i.e. the probability that each word in the vocabulary occurs in the evaluated context.

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i) \qquad (1)$$

In this work, we use entropy to assign a score to each gap based on the number of valid words that could fill in the slot given the surrounding context. As a result, gaps with many possible answers will yield higher entropy than those with fewer answers.

4 Experiments

We followed Malafeev's (2014) approach and used open cloze tests from Cambridge English examinations as our gold standard data, since they are manually created by experts in the field of EFL testing. We collected the sample open cloze tests for KET, FCE, CAE and CPE exams that are featured in their respective online handbooks[1] (one per exam together with their answers). These exams correspond respectively to levels A2, B2, C1 and C2 in the Common European Framework of Reference for Languages (CEFR). An open cloze test is not included in the PET (B1) exam, which is why it has not been included in our experiments.

For each exam, we restored the original text by using the answers provided (using the first alternative if there were many) and created 10 different variations of the open cloze tests by inserting gaps randomly throughout the text. We created the same number of gaps as in the original tests.

For each original and automatically generated test, we compute entropy per gap using a 5-gram language model trained on the 1 Billion Word WMT 2011 News Crawl corpus[2] using KenLM (Heafield, 2011). We use the language model bidirectionally, taking 3 words to the left and right of each gap to predict the probability of the next and previous words respectively. Since we obtain a probability for all the words in our vocabulary ($> 82,200$ words) given the left and right context individually, we multiply the probabilities for each word to get a unified "bidirectional" probability (see Figure 1). Given that this can lead to infinitesimal probabilities that can affect computa-

[1] https://www.cambridgeenglish.org/exams-and-tests/

[2] https://www.statmt.org/lm-benchmark/

Electronics firms, for example, expect to have only six months after they have introduced a new product before a rival company produces a _____ efficient or cheaper alternative.

⟶		⟵
wide	more	more
variety	very	energy
lot	less	is
quarter	is	as
line	and	less
...	...	...

Figure 1: An example calculation of candidate answers for a gap using the left and right context (in red). Candidate words are ranked from the most to the least probable.

tion, we use only the top 100 most probable words when computing entropy for each gap.

4.1 Results

Table 1 shows information about our gold standard tests, including CEFR levels, number of gaps and average gap entropy. The average gap entropy correlates positively with CEFR levels, suggesting that entropy increases with proficiency levels.

We then computed the average gap entropy for each of the 10 automatically generated tests per exam and compared them to the gold standard. Results are shown in Table 2.

Unlike the handcrafted gold standard, the automatically generated tests were produced randomly by a machine with no knowledge of test design so we would expect automatic gaps to be often inserted in inconvenient locations within the text, yielding lower quality tests. This hypothesis is verified by looking at the average gap entropy for the automatic tests, which is much higher than for the gold standard in the majority of cases (77.5%). This supports our intuition that entropy can be used to discriminate between good and bad gaps and, consequently, between good and bad tests.

We noticed that automatically generated tests for CPE tend to have lower entropy than the gold standard, contradicting our assumption in principle. However, we do not believe that these lower values indicate better tests but rather that they deviate from the expected difficulty for this proficiency level. In fact, we would expect high-quality tests to have average gap entropy around that of the gold standard tests, not too far below or over this reference value. Based on this premise, better automated tests can be constructed by controlling the entropy of gaps in the text, in line with previous work by Lee et al. (2019).

Exam	CEFR level	Number of gaps	Avg. gap entropy
KET	A2	11	1.29 ± 0.69
FCE	B2	9	2.33 ± 1.28
CAE	C1	9	2.69 ± 1.22
CPE	C2	9	5.16 ± 3.38

Table 1: Characterisation of our gold standard data.

4.2 Analysis

We looked at the gaps with the lowest and highest entropy to analyse how these values relate to the surrounding contexts. Table 3 shows the gaps in our gold standard tests with the lowest and highest entropy.

First, we found that gaps with the lowest entropy correspond mostly to exams at low CEFR levels while those with the highest entropy correspond to the highest CEFR level. This confirms our initial finding that entropy correlates directly with proficiency levels.

Second, we observed that gaps with low entropy are very restricted in context and built around very simple grammatical structures or vocabulary, making it easy to figure out the answers. On the other hand, gaps with high entropy are part of more complex grammatical structures and require longer context or understanding in order to be solved. This explains why our language model is unable to estimate the right answers for complex gaps, leading to higher entropy.

Finally, we investigated the correlation between entropy and the number of valid answers per gap. Pearson correlation for gaps in our gold standard tests is reported in Table 4. Contrary to our intuition, there is no consistent relationship between entropy and the number of valid answers per gap in our gold standard: KET shows negative correlation while CPE shows moderate positive correlation. We hypothesise that this is due to a limitation of the language model used in this preliminary study, which is unable to estimate the right word probabilities for gaps in complex contexts for the reasons described above. Using a more sophisticated language model should ameliorate this problem.

In any case, the values of entropy computed with our current model seem to capture the complexity of the gaps in context, which serves as a measure of difficulty. This, combined with the positive correlation with CEFR levels, makes en-

Exam	Average gap entropy per test									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
KET	5.40	3.47	3.63	3.73	4.22	4.18	4.60	4.33	4.74	4.34
FCE	4.53	7.13	6.12	4.30	**2.23**	4.01	2.45	3.93	4.18	5.67
CAE	4.70	4.66	3.57	**2.68**	3.26	4.38	2.79	5.08	5.07	4.82
CPE	6.58	**3.91**	**2.72**	**4.43**	**5.02**	**4.18**	5.83	5.46	**4.02**	**3.26**

Table 2: Average gap entropy for the automatically generated tests. Values lower than the gold standard are marked in bold.

Exam	Gap in context	Entropy	Answers
FCE	*..., apart _____ some minor mechanical problem...*	0.01	*from*
KET	*But _____ is some good news!*	0.23	*there / here*
KET	*_____ this okay?*	0.43	*is*
CPE	*... modern robots are dumb automatons, _____ of striking up relationships with their human operators.*	8.40	*incapable*
CPE	*Phones and computers have already shown the _____ to which people can develop relationships with...*	8.65	*extent / degree*
CPE	*Although sophisticated _____ to assemble cars...*	9.66	*enough*

Table 3: Example gaps with the lowest and highest entropy.

Exam	Pearson's ρ
KET	-0.1518
FCE	0.2333
CAE	0.0908
CPE	0.5149

Table 4: Correlation between entropy and the number of valid answers per gap.

tropy a suitable unsupervised evaluation measure for gaps in open cloze tests and encourages future work beyond this pilot study.

5 Conclusion and Future Work

This work investigated the use of entropy as an evaluation measure for gaps in open cloze EFL tests. Our study revealed that the average gap entropy of a test correlates positively with proficiency levels, so easier tests will contain gaps with lower entropy. A comparison between randomly generated tests and the handcrafted gold standard tests showed that the former had much higher entropy in general, confirming our intuition that generating random gaps is not optimal and that entropy can be used to discriminate between good and bad tests.

We also investigated the correlation between entropy and the number of valid answers per gap but results showed no consistent relationship, most likely due to the limitations of the n-gram language model used in this preliminary work. However, entropy was found to be a suitable proxy for gap complexity, which can be used to control the automatic generation of open cloze tests. Future work will address the limitations in this pilot study and investigate entropy on a larger sample.

References

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2014. Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics*, 2:517–529.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2015. Candidate evaluation strategies for improved difficulty prediction of language tests. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–11, Denver, Colorado. Association for Computational Linguistics.

Lisa Marina Beinborn. 2016. *Predicting and Manipulating the Difficulty of Text-Completion Exercises for Language Learning*. Ph.D. thesis, Fachbereich Informatik, Technische Universitt Darmstadt, Darmstadt, Germany. PhD thesis.

Jonathan C. Brown, Gwen A. Frishkoff, and Maxine Eskenazi. 2005. Automatic question generation for vocabulary assessment. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 819–826, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Ji-Ung Lee, Erik Schwan, and Christian M. Meyer. 2019. Manipulating the difficulty of c-tests. *CoRR*, abs/1906.06905.

John Lee and Stephanie Seneff. 2007. Automatic generation of cloze items for prepositions. In *Eighth Annual Conference of the International Speech Communication Association*, pages 2173–2176, Antwerp, Belgium.

Y. Lin, L. Sung, and M. Chen. 2007. An automatic multiple-choice question generation scheme for english adjective understanding. In *Workshop on Modeling, Management and Generation of Problems/Questions in eLearning*, pages 137–142. 15th International Conference on Computers in Education (ICCE 2007).

Alexey Malafeev. 2014. Language exercise generation: Emulating cambridge open cloze. *Int. J. Concept. Struct. Smart Appl.*, 2(2):20–35.

Edison Marrese-Taylor, Ai Nakajima, Yutaka Matsuo, and Ono Yuichi. 2018. Learning to automatically generate fill-in-the-blank quizzes. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 152–156, Melbourne, Australia. Association for Computational Linguistics.

Ruslan Mitkov and Le An Ha. 2003. Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing*, pages 17–22.

Irina Pandarova, Torben Schmidt, Johannes Hartig, Ahcène Boubekki, Roger Dale Jones, and Ulf Brefeld. 2019. Predicting the difficulty of exercise items for dynamic difficulty adaptation in adaptive language tutoring. *International Journal of Artificial Intelligence in Education*.

Juan Pino and Maxine Eskenazi. 2009. Measuring hint level in open cloze questions. In *Twenty-Second International FLAIRS Conference*, pages 460–465.

Juan Pino, Michael Heilman, and Maxine Eskenazi. 2008. A selection strategy to improve cloze question quality. *Intelligent Tutoring Systems for Ill-Defined Domains: Assessment and Feedback in Ill-Defined Domains.*, page 22.

Keisuke Sakaguchi, Yuki Arase, and Mamoru Komachi. 2013. Discriminative approach to fill-in-the-blank quiz generation for language learners. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–242, Sofia, Bulgaria. Association for Computational Linguistics.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Adam Skory and Maxine Eskenazi. 2010. Predicting cloze task quality for vocabulary training. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 49–56, Los Angeles, California. Association for Computational Linguistics.

Simon Smith, P. V. S. Avinesh, and Adam Kilgarriff. 2010. Gap-fill tests for language learners: Corpus-driven item generation. In *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing*, pages 1–6. Macmillan Publishers.

Eiichiro Sumita, Fumiaki Sugaya, and Seiichi Yamamoto. 2005. Measuring non-native speakers' proficiency of english by using a test with automatically-generated fill-in-the-blank questions. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, EdAppsNLP 05, pages 61–68, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yuni Susanti, Hitoshi Nishikawa, Takenobu Tokunaga, and Hiroyuki Obari. 2016. Item difficulty analysis of english vocabulary questions. In *CSEDU 2016 - Proceedings of the 8th International Conference on Computer Supported Education*, volume 1, pages 267–274.