A Corpus Study on the Difference between MOOCS and Real Classes

Adel Rahimi
Sharif University of Technology, Tehran, Iran
Tehran, Azadi Avenue, Iran
Tel.: +98 21 6601 3126
Adel.rahimi@mehr.sharif.edu

Parvaneh Khosravizadeh
Sharif University of Technology, Tehran, Iran
Tehran, Azadi Avenue, Iran
Tel.: +98 21 6601 3126
khosravizadeh@sharif.edu

Abstract

In this paper we take a look at how the language of Online classes (MOOCs) differs from those of real classes. Three corpora were created for this analysis; MOOC corpus, Lecture Capture corpus, and Philosophy Lecture Capture. Three factors were used in the study: Formality, Sentiment analysis, vocabulary analysis. Formality score was used to understand how formal the text is. Sentiment categorization of words was used to realize the positivity of the words used in the classes and finally top words used in corpora was analyzed to understand the usage. It was realized that the formality measure of real classes is slightly lower than online classes and professors use more positive words in real classes than online classes and the vocabulary usage is heavily under the influence of subject.

Keywords: MOOC, Online Classes, Sentiment analysis, computer-related courses

1. Introduction

MOOC stands for Massive Online Open Course. The first word, Massive, denotes the fact that MOOCs are "all-at-once-ness" (Johnson, Nafukho, Valentin, Le counte & Valentin, 2014). In other words, MOOCs are created once but will be distributed through internet platforms many times. Stephen Downes and George Siemens created the first MOOC in 2008 and used this term for the first time in a course titled "Connectivism and connective knowledge" (Downes, 2012, p.10). This particular class had 2,000 nonpaying students enrolled (Daniel, 2012).

In 2011 Stanford University offered the course Introduction to Artificial Intelligence. Initially 160,000 students enrolled and over 20,000 students completed the course. Udacity focuses on free education. Udacity incorporation, founded by Sebastien Thrun, was the first company which began to offer online courses. Since then, San Jose University, MIT, and Harvard among others began to offer on-line courses and establish MOOC platforms.

In terms of quality, the courses in MOOCs were just class materials published online by university professors, however, currently instructors are designing high quality tailored materials.

Different types of MOOCs have evolved from the beginning; EdX, Khan Academy and Coursera employ their very own style in creating class contents.

Table 1. Comparison of key aspects of MOOCs or Open Education initiatives from (Yuan, 2013)

#	First name and family name	For Profit	Free to access	Certification	Institutional Credits
1	EdX	×	✔	✔	×
2	Coursera	✔	✔	✔	✔
3	Udacity	✔	✔	✔	✔
4	Udemy	✔	✔	✔	✔
5	P2PU	×	✔	×	×

Based on class-central.com, as of 2016 there are more than 4,000 MOOCs.
Distribution of MOOCs over subjects (data by class-central.com):

Table 2. Comparison of MOOCs distribution in different subjects (data from: class-central.com)

Subject	**Percent**
Science	11.3%
Business & Management	16.8%
Mathematics	4.09%
Engineering	6.11%
Art and Design	6.73%
Programming	7.44%
Health and Medicine	8.27%
Education and Teaching	9.36%
Humanities	9.41%
Computer Science	9.74%
Social Sciences	10.8%

As Belanger and Thornton (2013) suggest, the main reasons behind the popularity of MOOCs are;

• To support lifelong learning or gain an understanding of the subject matter, with no particular expectations for completion or achievement,

• For fun, entertainment, social experience and intellectual stimulation,

• Convenience, often in conjunction with barriers to traditional education options,

• To experience or explore online education.

2. Related works

Although MOOCs have a brief history, they have evolved so vastly during present time. Several studies have been conducted on how MOOCs were started. Daneil (2012) describes a short history of MOOCs and expands it in the wider context of distant learning. Yuan (2013) describes a history of MOOCs and an analysis on the MOOC-style open educations. Also describes the challenges for MOOCs. Clow (2013) in a paper titled "MOOCs and the funnel of Participation" uses funnel as a metaphor for describing dropout rates in MOOCs. Jordan (2014) used public dataset to visualize the completion rates on MOOCs.

In the terms of being difficulty level Konnikova, (2014)[1] states that if MOOCs were to challenge students they would likely be more effective.

Keats, (2016)[2] in an article in wired magazine describes a history of MOOCs and writes that MOOCs should be expansive in order to be successful and replace formal education. In another paper Dellarocas and Alstyne (2013) explains business models for MOOCs and how making money out of MOOCs would work.

Chen (2014) used text mining to understand the challenges of MOOCs. His study showed that among other challenges, MOOCs need to overcome course quality, high dropout rates, unavailable course credits, ineffective assessments, and complex copyright issues.

Rodriguez (2012) classifies MOOCs into two categories: AI-Stanford and connectivist MOOCs (c-MOOCs). Rodriguez (2012) suggests that c-MOOCs are more social than AI-Stanford.

Jordan (2014) also reported completion data on 24 MOOCs the data shows the highest competition rate was for Functional Programming Principles in Scala which was 19.2%.

[1] Konnikova, M (2014). Will MOOCs be flukes? The New Yorker, Retrieved on July, 21, 2017 from www.newyorker.com/science/mariakonnikova/moocs-failure-solutions

[2] Keats, J. (2016). Are MOOCs in danger of becoming irrelevant? The New Yorker. Retrieved on August, 10, 2017 from http://www.wired.co.uk/article/improving-moocs-jonathon-keats

Jordan (2014) also reported that most MOOCs had 43,000 students enrolled but the completion rate is only 6.5%.

Although many students drop out from the course, Onah, Sinclair, and Boyatt (2014) shows that many participants follow the course in their own "preferred way". Onah et al. (2014) also suggests "structure of 'a course' may not be helpful to all participants and supporting different patterns of engagement and presentation of material may be beneficial."

Reasons for dropout suggested by Onah et al. (2014) are: No real intention to complete, Lack of time, Course difficulty and lack of support, lack of digital skills or learning skills, bad experience, bad expectations of the course, starting late, peer review.

Brinton, Chiang, Jain, Lam, Liu, and Wong (2014) analyzed discussion forums in the MOOCs and identified two features of the discussion forums in the MOOCs: 1) high decline rate, 2) high volume noisy discussions. Brinton et al. then proposes a unified generative model for discussion threads and an algorithm for "Ranking thread relevance".

Wen, Yang, and Rose (2014) uses sentiment analysis to "monitor students' trending opinions towards the course and major course tools". Wen et al. (2014) also reported that there is a high correlation between number of dropouts and sentiments expressed in the discussion forums.

3. MOOC/LC Corpus Design

The main subject of this study is focused towards Computer Science and Computer related courses.

The following 3 corpora were prepared for this study:

1. MOOCs Computer Corpus (Computer)
2. Lecture Capture Corpus (Computer)
3. Lecture capture (Philosophy)

Table 3. List of courses used in the MOOC corpus

Course	**Course provider**	**Number of Word Token**
Computer science 101	Stanford University	79039
Natural Language Processing	Columbia University	136215
An Introduction to Interactive Programming in Python (Part I)	Rice University	145336
Text Retrieval and Search engines	University of Illinois at Urbana-Champaign	77986
Neural Networks	University of Toronto	122714
Digital Signal Processing	École Polytechnique Fédérale de Lausanne	154333
CS1 Compilers	Stanford University	196639
Computational Investing, Part I	Georgia Institute of Technology	74173
C++ for c programmers	University of California, Santa Cruz	63327
Biology Meets programming: bio-informatics for beginners	University of California, San Diego	10257
Audio Signal Processing for Music Applications	Stanford University	144364
		1204383

The corpus is made of Coursera class subtitle which is exactly what the instructor is saying. Some files were originally in .srt format. The subtitle was first converted to .txt then a cleaning method was applied on the corpus meaning all the numbers and special characters were removed.

The second corpus is from real university classes referred to as Lecture Capture. This corpus is the data from MIT OCW, CS50 website, and other course websites that offer closed captions for people with hard of hearing.

Table 4. List of courses used in the LC Corpus

Course	**Course Provider**	**Number of word tokens**
6.001	MIT	121245
CS50	Harvard University	342458
Computer Science E-76: Building Mobile Applications	Harvard University	133298
Computer Science E1 Computers and Internet	Harvard University	223938
CS50 (2016)	Harvard University	246036
Software Engineering	Harvard University	66783
		1133758

The Third corpus is the lecture capture corpus for philosophy classes. Same as lecture capture corpus the corpus is compiled from course websites.

Table 5. List of courses used in the Philosophy Lecture Capture corpus

Course	**Course Provider**	**Number of word tokens**
Philosophy and the Science of Human Nature	Yale University	159210
Introduction to Political Philosophy	Yale University	145114
Death	Yale University	191559
		495883

The reason behind choosing several courses in MOOCs and only few courses from LC corpus is that the length of the classes in universities is much longer than those of the MOOCs and therefore to stratify the corpus, the number of MOOCs is higher.

In order for the corpora to be in the same category, Courses have been chosen meticulously so that: Firstly, they do not differ in terms of theme, class organization, and other possible affecting factors. Secondly, the number of word tokens in the corpora to be roughly equal so that it does not affect quantitative factors.

4. Analysis

4.1. Formality

To analyze formality in the corpus, (Heylighen and Dewaele, 1999) F-score measurement was employed to indicate how formal the instructors' speech is. As the formula supposes:

F = (noun frequency + adjective freq. + preposition freq. + article freq. – pronoun freq. – Verb freq. – adverb freq. – interjection freq. + 100) / 2

Thus, initially it is required to analyze parts of speech in the corpora. In the first step Stanford POS tagger was used to tag all the words in both corpora.

MOOC Corpus F-Score:

F = (19.466+ 1.635+ 7.144+ 3.267+ 10.456+ 12.6230- 6.09- 18.010- 7.144- 3.521+100) / 2 = **59.9**

The F-Score in MOOCs is slightly higher than the other two. This 10 percent doesn't drastically change the formality whereas; it can be a measure for further studies on the formality of online classes.

Lecture Capture Corpus F-Score:

F = (1.132+19.66+6.626+0.12+3.38+10.57-6.56-18.676- 8.874-4.04+100) / 2 = **51.67**

Philosophy LC corpus F-Score:

F = (9.2 + 0.53 + 3.48+ 4.94+ 1.44+ 4.420- 3.03- 7.76- 2.611- 1.348+ 100) / 2 = **54.63**

The formality score for Philosophy Lecture Capture is ~55%. This number is slightly higher than the Lecture capture. This shows that the subject has a role in the formality of class as well as the course platform.

4.2. Sentiment

Sentiment analysis can be employed as a measure to study how positive or negative the lecturers' speech actually is. In this measure, the AFFIN wordlist Nielsen (2011) was used. The AFFIN wordlist is a list of vocabularies based on positive or negative sentiment of each word.

Table 5. Distribution of words by sentiment in the MOOC corpus

	Frequency	Percent
-4	16	.0
-3	804	2.2
-2	5650	15.4
-1	4591	12.5
1	9112	24.9
2	13306	36.3
3	2770	7.6
4	357	1.0
5	2	.0

More than 69 percent of the words used in the MOOC classes, in the computer-related subjects, are positive words. The most frequent positive category is +2 and the most frequent non-positive word category is -2.

Table 6. Distribution of words by sentiment in the Lecture Capture corpus

	Frequency	Percent
-4	3	.0
-3	131	.3
-2	1705	3.3
-1	6499	12.5
1	6077	11.7
2	12938	24.9
3	19541	37.6
4	4502	8.7
5	637	1.2

In the Philosophy lecture capture class, percentage of +3 words is the highest. Also the most frequent negative word is -1.

Table 7. Distribution of words by sentiment in the Lecture Capture corpus

	Frequency	Percent
-4	1	.0
-3	32	.3
-2	901	9.7
-1	1930	20.7
1	972	10.4
2	1909	20.5
3	2675	28.7
4	825	8.8
5	85	.9

The lecture capture is slightly different in terms of category distribution. Approximately 50% of the corpus is in the positive category.

Table 8. Distribution of words by sentiment in the all corpora

	MOOC	LC	Philosophy Lectures
-4	.0	.0	.0
-3	2.2	.3	.3
-2	15.4	3.3	9.7
-1	12.5	12.5	20.7
1	24.9	11.7	10.4
2	36.3	24.9	20.5
3	7.6	37.6	28.7
4	1.0	8.7	8.8
5	.0	1.2	.9
Positive	69.8	84.1	69.3

The above Table shows that the corpus with the most frequent positive vocabulary, having more than 80% of its words in the positive category, is the LC corpus. The philosophy LC and the MOOC classes were more or less the same having ~70%.

4.3. Top Vocabularies

- MOOCs:

going
one
let
okay
see
two
like
use
get

- LC corpus top vocabularies:

going
like
one

let
go
right
actually
see
get
want

- Philosophy LC corpus top vocabularies:

think
say
would
us
life
question
going
things
way
like

In the terms of top vocabulary, difference between MOOC and LC is little but the Philosophy LC is much more different than the other two. This indicates that choosing frequent vocabularies is under the influence of the subject.

5. Results

The F score shows the MOOCs in general, are slightly more formal, and in particular computer-related courses (referred to as MOOC corpus), is the most formal among all. Sentiment distribution shows that lecture capture is more positive in term of word usage, while MOOC in the second place, and philosophy LC in the third, suggest subject might have an impact on the word usage. The top vocabulary list states that, subject heavily influences the frequent words instructors choose, and thus it will not change in real or virtual classes.

These data show how and why MOOCs are different from real classes and how instructors can get the most out of their class time.

References

Belanger, V., Thornton, J. (2013). *Bioelectricity: A Quantitative Approach*. Duke University Press.

Brinton, C. G., Chiang, M., Jain, S., Lam, H., Liu, Z. & Wong, F. M. F. (2014). Learning about social learning in MOOCs: From statistical analysis to generative model, *IEEE Transactions on Learning Technologies*, *7*(4), pp. 346-359.

Chen, Y. (2014). Investigating MOOCs through blog mining. *The International Review of Research in Open and Distributed Learning*, *15*(2), pp. 85-106.

Clow, D. (2013). MOOCs and the funnel of participation. In Proceedings of *the Third International Conference on Learning Analytics and Knowledge*. New York, NY, USA: ACM. pp. 185-189.

Daniel, J. (2012). Making sense of MOOCs: Musings in a Maze of myth, paradox and possibility. *Journal of Interactive Media in Education.* 2012(3). DOI: http://doi.org/10.5334/2012-18

Dellarocas, C., & Van Alstyne, M. (2013). Money models for MOOCs. *Communications of the ACM, 56*(8), pp. 25-28.

Downes, S. (2012). *Connectivism and connective knowledge: Essays on meaning and learning networks*. Retrieved from http://www.downes.ca/files/books/Connective_Knowledge-19May2012.pdf

Heylighen, F., & Dewaele, J. M. (1999). Formality of language: Definition, measurement and behavioral determinants. Internal Report, Center "Leo Apostel", Free University of Brussels.

Johnson, D., Nafukho, F., Valentin, M., Lecounte, J., & Valentin, C. (2014). The origins of MOOCs: The beginning of the revolution of all at once-ness. In Proceedings of *15th International Conference on Human Resource Development research and practice across Europe*. Edinburgh, UK: Edinburgh Napier University Business School.

Jordan, K. (2014). Initial trends in enrolment and completion of massive open online courses. *The International Review of Research in Open and Distance Learning, 15*(1), 133-160.

Nielsen, F. (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs. In Proceedings of *the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*. Heraklion, Crete, Greece, May 30, 2011

Onah, D. F. O., Sinclair, J. and Boyatt, R. (2014). Dropout rates of massive open online courses: Behavioural patterns. In proceedings of the *6th International Conference on Education and New Learning Technologies*, Barcelona, Spain, 7-9 Jul 2014. pp. 5825-5834.

Rodriguez, C. O. (2012). MOOCs and the AI-Stanford like courses: Two successful and distinct course formats for massive open online courses. *European Journal of Open, Distance and E-Learning, 15*(2).

Wen, M., Yang, D. & Rose. C. P. (2014). Sentiment analysis in MOOC discussion forums: What does it tell us? In Proceedings of the *7th International Conference on Educational Data Mining (EDM 2014)*, pp. 130–137, 2014.

Yuan, L., & Powell, S. (2013). MOOCs and open education: Implications for higher education. *Centre for Educational Technology & Inoperability Standards*. Retrieved from http://publications.cetis.ac.uk/wp-content/uploads/2013/03/MOOCs-and-Open-Education.pdf.

Adel RAHIMI is MSc student of computational linguistics at Sharif University of Technology. His research interests include: Machine Learning, Natural Language Processing, and Computational Linguistics. He is currently member of the Sharif Speech and Language Processing lab.
You can visit the personal website of Mr Adel Rahimi at: http://adelr.ir/

Parvaneh KHOSRAVIZADEH is an assistant professor at Sharif University of Technology in the field of computational linguistics. Her research interests include Machine Learning, Machine Translation, Discourse analysis, and Psycholinguistics. You can see the Google Academic profile of Professor Khosravizadeh here: https://scholar.google.com/citations?user=UcH_97cAAAAJ&hl=en