Lessons from a Massive Open Online Course (MOOC) on Natural Language Processing for Digital Humanities

Simon Clematide, Isabel Meraner, Noah Bubenhofer, Martin Volk
University of Zurich, Switzerland
Institute of Computational Linguistics
simon.clematide@uzh.ch, isabel.meraner@uzh.ch,
bubenhofer@cl.uzh.ch, volk@cl.uzh.ch

Abstract

In this paper, we present the concept, content and experience with an actively running Massive Open Online Course (MOOC) on Natural Language Processing for Digital Humanities. This video-based course is held in German, does not require any programming skills, and serves as an introduction to automatic text analysis. The target audience is anyone who is interested in applying basic language technology to text corpora. It has a strong empirical focus on digital representations, tools and corpus linguistics. The main goal thereby is to grasp the fundamental terminology and concepts of computational linguistics, to understand the main problems and solutions, as well as to know about the performance and limitations of current methods. Furthermore, manual annotation and data visualization are introduced in this course.

1 Introduction

More and more scientific disciplines use automatic text analysis in their digital scholarship. In the humanities, we have literary and cultural studies (e.g. popularized as "distant reading" (Moretti, 2013), "corpus based discourse analysis" (Sinclair, 2004; Bubenhofer, 2009) etc.), empirical corpus linguistics and computational social sciences (including automatic media monitoring (Reamy, 2016)), but text mining is also popular in the natural sciences, for instance in the bio-medical domain (Cohen and Hunter, 2008).

Being able to apply Natural Language Processing (NLP) methods to texts requires special knowledge and skills. The goal of this course is not to teach these skills, but to didactically introduce important concepts and techniques related to digital text representation and analysis. Therefore, programming experience is neither required for this introductory course nor provided in it.

According to Ubell (2017), more than 58 million people have signed up worldwide for Massive Open Online Courses (MOOCs) by now. This form of distance learning in higher education has grown popular over the last 6 years and several commercial and non-commercial platforms compete for participants.

Our free course is held on Coursera[1], one of the largest commercial platforms that distributes classes mostly held in English and created by lecturers of top universities around the world. Our course language is German[2] which on the one hand has the disadvantage of excluding participants who do not speak German, but on the other hand, it allows us to occupy a niche in language technology focusing on German texts. A first session of the course was run in summer 2015, and about 900 learners visited the course at least once. Due to legal issues between our university and Coursera, and due to the introduction of Coursera's new platform[3] and the resulting course migration effort, it took two years to start the next session of our course.

The rest of this paper is organized as follows: in section 2, we introduce and motivate the syllabus of our course, in section 3 we discuss our experience from running the course twice so far.

[1] The MOOC can be accessed via this link: https://www.coursera.org/learn/digital-humanities.

[2] All videos have German subtitles, which is especially useful for users with a limited understanding of German. We explicitly allow English contributions in the discussion forum and peer assignments.

[3] Coursera now offers all courses more flexibly on demand by restarting each course regularly at intervals of several weeks. Learners can now easily switch from one instance of a course to the next if they cannot complete within their initial learner cohort. According to Saraf (2017), these cohorts improve the completion rate compared to purely self-paced learning and still offer more flexibility.

2 Course Structure

The course is designed to run over a period of 6 weeks each of which has its own thematic focus. Each thematic module consists of 30 to 90 minutes of videos, which mostly use a fairly traditional format where a lecturer presents slides and explains NLP methods in an accessible and illustrative way. In addition, we provide learner-oriented learning objectives, more detailed readings regarding the presented topics and further course material within each module. In order to test the individual learning progress, we integrated either a brief final quiz or a peer assessment at the end of each module. The course syllabus is structured as follows:

Module 1 "Paths into the Digital World"
This introductory module presents the fundamental concepts and terminology regarding the digitization of texts. We present techniques such as scanning and OCR (Optical Character Recognition) as well as other approaches for the acquisition of text corpus material (including digital-born documents), and we discuss potential problems related to digitization and corpus design. Additionally, short interviews about digitization techniques and the relevancy of digitization with two experts from the Zurich central library complete the first module.

Module 2 "Structured and Effective Representation of Corpus Data"
The second module provides an overview of different encodings, the markup language XML and the TEI P5 standard for text representation. The second half of the module has its focus on automatic tokenization and sentence segmentation. Finally, in a non-graded hands-on discussion prompt the learner needs to apply the acquired XML knowledge concerning well-formedness and identify syntax errors in an XML document.

Module 3 "Properties of Corpora and Basic Methods for Analysis"
In this module, we present the basic concepts of corpus linguistics such as term frequencies, n-grams, collocations and methods for analyzing texts according to Lemnitzer and Zinsmeister (2006). In addition, we demonstrate the functionality of various platforms and interfaces for corpus analysis and show some hands-on corpus query examples. In the last part of Module 3, we introduce the topic "visual linguistics" (Bubenhofer, 2016) together with a variety of tools for displaying the properties of texts in a creative, interactive and illustrative way.

Module 4 "Automatic Corpus Annotation Using NLP Tools"
In this module, we introduce different automatic corpus annotation methods, such as part-of-speech tagging, lemmatization, stemming, parsing, Named Entity Recognition, and Entity Linking (Ratinov and Roth, 2009) for automatic disambiguation. Furthermore, we investigate potential problems and sources of errors that can emerge while using such automatic annotation tools and we offer approaches to solving these issues.

Module 5 "Manual Annotation and Evaluation of Corpus Data"
The main topic of module 5 is the efficient combination between manual and automatic annotation and the integration of machine learning methods in the vein of Pustejovsky and Stubbs (2013). Subsequently, we present the most common metrics for measuring the quality of NLP systems and introduce the concept of inter-rater reliability. In the second part of module 5, we focus on the possibilities and restrictions of crowd-sourcing methods in the digital humanities.

Module 6 "Challenges in Multilingual Text Analysis"
The last module concentrates on multilingual and parallel corpora as well as on automatic language identification in large-scale text collections. Finally, we introduce several up-to-date tools for automatic alignment of parallel corpora on the level of documents, sentences and words.

Assessments
In terms of graded assignments, we integrate short single and multiple choice quizzes ranging from 5 to 12 questions at the end of each module[4]. Module 3 and Module 5 additionally include a graded peer assignment where each learner is supposed to assess at least two other submissions according to detailed grading instructions. The peer assessment in Module 3 encourages the learner to apply the acquired knowledge on complex corpus queries. By this means, each learner performs individual queries on the IMS Open Corpus Workbench or the COSMAS II interface, regarding diachronic language change. Apart from this, the learner is supposed to generate frequency charts or collocation

[4] One module currently does not have a quiz.

profiles and to interpret the findings and insights gained from this task.

The peer assignment in module 5 demands the learner to run the online demo version of the Stanford Named Entity Tagger (Finkel et al. (2005)) or the Thomson Reuters Open Calais (Reuters (2008)) on a small sample text of his own choice and to evaluate the NER taggers's output according to the evaluation metrics precision and recall that we explained in this module. In this manner, peer assessments motivate the learners to try out individually different NLP tools and corpus query platforms, and to question and critically analyze their output.

Community building and feedback

In order to enhance community building and thematic exchange between enrolled learners, we included a "Meet and Greet" discussion prompt section in the first module as well as a "Feedback and Thank You" discussion field at the very end of the last module. For each module a weekly discussion forum is automatically generated on the platform where participants can ask or answer questions regarding the content of a module. Additionally, for each discussion prompt, individual threads are automatically included in the weekly forum to allow topic-related discussions and exchange. To ensure a friendly discussion atmosphere and make the learners feel well looked after, the course tutor is actively present in the forums and tries to answer or comment every contribution.

2.1 Lecturers and tutors

Three different lecturers teach in this course and they agreed beforehand on the overall content, syllabus and presentation style. After that, each lecturer was responsible for developing his own module content, preparing the slides and organizing additional material. A student assistant supported this process, cut the video recordings, added some video effects (zooming, highlighting, textual annotations, in-video quizzes in order to avoid monotony) to the slide recordings and published everything on Coursera's electronic learning management platform. All lecturers already had a lot of teaching experience in the subjects of their modules, yet, everyone had to invest a large amount of time to fit the existing teaching material from normal university classes into video sequences of an appropriate length for online courses. Actually, some of our videos are still too long by current standards (5-7 minutes).

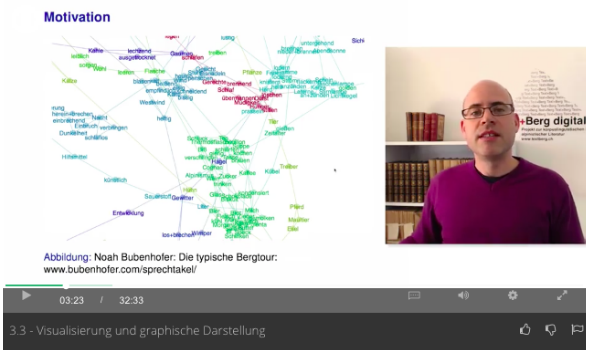
Motivation
Abbildung: Noah Bubenhofer: Die typische Bergtour:
www.bubenhofer.com/sprechtakel/
+Berg digital
03:23 / 32:33
3.3 - Visualisierung und graphische Darstellung

Figure 1: Talking head recording of speaker presenting slides in a weekly video session.

Having 3 different lecturers makes the course more varied and offers the learner slightly different perspectives on the matter. In addition, every lecturer was able to teach the topics he is more specialized and experienced in.

2.2 Building a studio and gaining recording experience

For the time of the video recordings we turned an office into a makeshift studio. We decided to record the videos on our own and not by a multimedia production team from our university, who, however, instructed us kindly in the beginning. Although the result would have looked more professional, this gave us the urgently needed flexibility in production as all lecturers had no prior experience in teaching in front of a camera.

The scene background was white with some books and logos for ease of recognition (see Fig. 1). Lighting was installed to keep the scene equally illuminated without making the lecturers look pale. The lecturers were filmed from the side while sitting in order to offer a relaxed learning atmosphere. Lecturers needed a while to learn to keep eye-contact with the camera rather than looking at their slides. Small slips of the tongue were accepted as ingredients of natural talks. Larger wording errors were cut out and required repeated recordings.

2.3 Resources and NLP Tools

As a running example, we use the diachronic and multilingual corpus Text+Berg (Volk et al., 2010) which allows us to illustrate many different NLP tasks and exploitation techniques on a coherent and academically freely available resource. This corpus has texts mostly in French, German, and Italian, some of them translated, and spans over a period of 150 years.

In our videos, we also mention, demonstrate and reference a lot of other initiatives, resources, frameworks, and open-source tools: (a) digitization initiatives (Projekt Gutenberg, Europeana, TextGrid); (b) OCR crowd-correction and crowd-sourcing in general (TypeWright, Crowdflower, Artigo); (c) online corpora and corpus query tools (COSMAS II/DeReKo, DWDS, CQPweb); (d) parallel corpora (EuroParl, Canadian Hansard); (e) sentence and word alignment tools for parallel corpora (InterText, HunAlign, GIZA++); (f) language identification (lingua-ident, LangId); (g) text representation standards (Unicode, UTF-8, XML, TEI-P5); (h) annotation standards (STTS, Universal tags and dependencies); (i) standard lexical and syntactic NLP tools (Porter Stemmer, Durm Lemmatizer, TreeTagger, Connexor-Tagger; chunkers and parsers); (j) named entity recognition (Open Calais, Stanford NER); (k) tools for manual annotation of linguistic structures (and/or querying the annotations) (WebAnno, ANNIS, EXMARaLDA, RSTTool); (l) visualization (Graphviz, Leaflet, Gephi).

3 Discussion

The field of language technology and NLP is a rapidly evolving discipline. In the last 25 years, systems based on hand-written rules and application-specific algorithms have been largely superseded by statistical systems that are typically built by supervised or semi-supervised machine learning techniques.

In our course we reflect this paradigm change, e.g. by contrasting the output of a rule-based part-of-speech tagger with a statistical one, and make our participants aware of the different requirements for these approaches (e.g. manually built training material needed for supervised machine learning). However, we do not introduce "Neural Deep Learning" methods (Manning, 2015), which currently dominate NLP research and already have an impact on practical NLP systems. Our course design, which roughly follows the traditional NLP pipeline steps with language identification, tokenization, part-of-speech tagging, syntactic analysis and semantic analysis does not particularly fit the recent trend for neural end-to-end systems (Zhang et al., 2015), which – in the extreme – try to avoid these steps altogether and favor purely character-based approaches.

For an introductory course targeting the basics of text analysis for digital humanities and addressing learners with a mostly arts and humanities background, we strongly believe that our course structure results in a better understanding of the problems that one needs to tackle when processing natural language.

In addition, white box instead of black box systems using valid features[5] are most important for digital humanities and linguistics: It is often crucial to properly design linguistic meaningful features to receive valid categories for understanding the specificity of a text corpus or a linguistic phenomenon. To give a simple example: Even if a statistical model based on character n-grams turns out to perform best for authorship attribution, this model is of low interest for a linguistic research question on writing styles. That is because character n-grams do not represent a linguistic meaningful category and it is unclear what a character n-gram measures.

Even though: A follow-up intermediate course clearly would need to focus more on distributional (word embeddings and topic modeling) and neural approaches, which, however, require more knowledge in mathematics and programming skills.

Active Learning

Successful MOOCs have to offer more than just recorded video streams of lectures. Freeman et al. (2014) show that active learning settings generally improve the learning outcome of participants. Platforms such as Coursera offer several technical solutions for making distance learning more than passive consumption of videos. Individual user activity for an active and enduring learning experience is encouraged through several course items. In-video-questions re-captivate the learner's attention and require brief reflections on recently learned course content. Peer assignments encourage learners to apply knowledge from the current module and to try out NLP tools individually and critically evaluate their actual performance. By assessing other peers, further reflection and critical feedback is demanded from the learners. In a hands-on video in Module 3, we provide step-by-step instructions for individual corpus analysis with the IMS Open Corpus Workbench. A brief introduction of the CQP query language allows the learner to issue more complex queries. Furthermore, we constantly invite the learner to apply his or her newly acquired skills and do further experiments on his or her own at the end of each module.

[5] Valid for categories in the respective discipline.

Community building and forum activity

From the experience with the first session of the course held in summer 2015, there is only a limited need of the users for exchange in the forums. There was some discussion on more advanced topics such as dependency parsing which was mentioned in Module 3, however, more formally introduced only later in Module 4. In the past, the peer assessments on the evaluation of named entity taggers triggered some discussions, for instance, on the question whether the German word "Mittelmeerraum" (Mediterranean) should be recognized as a toponym or not.

Our course participants on Coursera come from all over the world[6], although naturally participants from the German speaking countries dominate. The participants have different backgrounds and interests, in our current course 37% declare themselves as higher education students. Others are either looking for a job after graduation or already employed and willing to expand their knowledge regarding NLP for Digital Humanities.

Course development

After successfully running our MOOC on the new Coursera learning management system in Summer 2017, we fine-tuned our course for its future iterations. We tried to respond to previous learner's feedback and to include a variety of small adjustments such as smaller quizzes after each video instead of longer quizzes at the end of each module. We now provide guidelines at the beginning of the MOOC and explain how the course can be used in order to satisfy wide-ranging needs of learners with different backgrounds, therefore easing "cherry-picking" of certain course modules and not forcing everybody into following the one-module-per-week order. Additionally, we integrated more discussion and reading prompts related to the course content to maintain the learner's active attention. A new outlook section in the last module provides further links and information on machine translation and recent trends on applying Neural Network methods in NLP. In October 2017, a new version of the course goes live where learners will be able to purchase a certificate provided by the platform Coursera that can be helpful when seeking a job in the field of Digital Humanities.

4 Conclusion

This paper presents the content of an ongoing introductory MOOC on Natural Language Processing for Digital Humanities. Any participant who successfully completes this course will have a broad overview on the problems and solutions for automatically enriching and exploiting text corpora (via visual exploration or more sophisticated corpus queries). The course introduces the process of digitization, corpus creation, text representation, statistical analysis, visualization, automatic and manual annotation on different linguistic levels, as well as the challenges and benefits of multilingual resources.

As with any MOOC, the number of participants that actually complete the course is only a small fraction (5 to 12%) of all registered users (Ubell, 2017). When our course was run for the first time in 2015, 46 participants achieved a certificate of accomplishment out of 883 learners who actually visited the course at least once. In the current on-demand setup of the course that started in July 2017 we have a lower number of registered learners, however, the majority of them seems to be actively following the course.[7]

The number of participants cannot be considered "massive" in the literal sense of "Massive Open Online Course", however, MOOCs actually do not need to have thousands of students. The strength of courses like ours lies in their openness, in the way they present and offer specialist knowledge to interested people all over the world, and last but not least, how they structure the learning process and the topics in an accessible way and easily digestible portions.

Acknowledgment

The production of our MOOC was financed by the division "Digitale Lehre und Forschung (DLF)" from the Faculty of Arts of the University of Zurich (UZH). We would like to thank Anita Holdener (DLF) for her constant technical support, Lukas Meyer from "Multimedia & E-Learning-Services (MELS)" of the UZH for producing our promotion video and for the introduction in video recording he gave to us, and last but not least, Sara Wick, our initiative student tutor and production assistant.

[6] Some of them are also motivated by the fact that the course is given in German.

[7] Regarding the participation in the course, we currently have 211 active learner out of 293 enrolled learners.

References

Noah Bubenhofer. 2009. *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. Sprache und Wissen, 4. De Gruyter, Berlin, New York.

Noah Bubenhofer. 2016. Drei Thesen zu Visualisierungspraktiken in den Digital Humanities. *Rechtsgeschichte Legal History - Journal of the Max Planck Institute for European Legal History*, (24):351–355.

K. Bretonnel Cohen and Lawrence Hunter. 2008. Getting started in text mining. *PLOS Computational Biology*, 4(1):1–3.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, 6:363–370.

Scott Freeman, Sarah L. Eddy, Miles McDonough, Michelle K. Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. 2014. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23):8410–8415.

Lothar Lemnitzer and Heike Zinsmeister. 2006. *Korpuslinguistik. Eine Einführung*. Narr, Tübingen.

Christopher D. Manning. 2015. Last words: Computational linguistics and deep learning. *Computational Linguistics*, 41:701–707.

Franco Moretti. 2013. *Distant Reading*. Verso Books, London.

James Pustejovsky and Amber Stubbs. 2013. *Natural language annotation for machine learning*. O'Reilly Media, Sebastopol, CA.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. *CoNLL*, 6:147–155.

Tom Reamy. 2016. *Deep text: using text analytics to conquer information overload, get real value from social media, and add big(ger) text to big data*. Information Today.

Thomson Reuters. 2008. Open calais demo. `http://www.opencalais.com/opencalais-demo/`. Date accessed: 20/07/2017.

Kapeesh Saraf. 2017. Life gets in the way: How Coursera is solving for the biggest challenge in online learning. `https://blog.coursera.org/life-gets-way-coursera-solving-biggest-challenge-online-learning/`. Date accessed: 20/07/2017.

John Sinclair. 2004. *Trust the Text. Language, Corpus and Discourse*. Routledge, London.

Robert Ubell. 2017. MOOCs come back to earth. *IEEE Spectrum*, 54(3):22–22.

Martin Volk, Noah Bubenhofer, Adrian Althaus, Maya Bangerter, Lenz Furrer, and Beni Ruef. 2010. Challenges in building a multilingual alpine heritage corpus. *Seventh International Conference on Language Resources and Evaluation (LREC)*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28*, pages 649–657. Curran Associates, Inc.