Detailed Description of the Development of a MOOC in the Topic of Statistical Machine Translation

Marta R. Costa-jussà, Lluís Formiga, Jordi Petit*, and José A.R. Fonollosa

Departament de Teoria de Senyal i Comunicacions
Departament de Ciències de la Computació
Universitat Politècnica de Catalunya, 08034 Barcelona
{marta.ruiz,lluis.formiga,jose.fonollosa}@upc.edu, jpetit@cs.upc.edu

Abstract. This paper describes the design, development and execution of a MOOC entitled “Approaches to Machine Translation: rule-based, statistical and hybrid”. The course is launched from the Canvas platform used by recognized European universities. The course contains video-lecture, quizzes and laboratory assignments. Evaluation is done using a virtual learning environment for computer programming and peer-to-peer strategies. This MOOC allows to introduce people from various areas to the Machine Translation theory and practice. It also allows to internationalize different tools developed at the Universitat Politècnica de Catalunya.

1 Introduction

The Massive Open On-line Courses (MOOCs) have received large acceptance among the academics to develop their courses. Among the different advantages, MOOCs require low resources in comparison to the international impact that they potentially have.

In the context of Natural Language Processing (NLP), there is the discipline of Machine Translation (MT) that puts together different communities from linguists to engineers. From the Universitat Politècnica de Catalunya, with more than 12 years of experience in the topic, we have launched a MOOC that overviews the basis behind the most popular MT systems. The main motivations to develop this course are listed as follows:

1. To improve the contents and digital resources of MT.
2. To adapt an MT course to an international public.
3. To propose the course in complete platform that is used by large European universities.
4. To use and publisize the open-source NLP tools developed at the UPC that are useful in the MT community.

This rest of this decriptive paper is organized as follows. Section 2 presents the course topic and contents. Next section details the activities proposed to the students. Section 4 describes the procedure to record the videolectures and the promotional video and section 5 explains which are the functionalities of the Canvas platform. Finally, section 7 concludes the paper with final remarks.

* Corresponding author.

© Springer International Publishing Switzerland 2014
2 Course Topic and Contents

The course aims at giving a broad view of MT. Given that MT is a multidisciplinary field involving linguists, computer scientists, engineers and informatics, this course is targeted for all these profiles.

The course will review the most important paradigms of MT including statistical, rule-based and hybrid translation, focusing on giving an overview of the state-of-the-art techniques and algorithms and outline the most challenging problems in the field. In addition, an overview of several evaluation approaches will be presented: automatic measures, human perception evaluation, and human linguistic evaluation, putting emphasis on the advantages and disadvantages of each method. The course will underline how statistics and linguistics are combined in both development and evaluation.

From the theoretical point of view, at the end of the course the student will understand the theory behind rulebased, statistical and hybrid MT systems. The student will have to pass several theoretical tests to show his skills. From the practical point of view, at the end of the course the student will be able to build a rule and a statistical-based MT system. The student will be asked to build parts of the systems along the course.

References that are pillars of the course include the Statistical Machine Translation book [1] and the survey of Linguistics applied to Statistical Machine Translation [2]. Other references have been the slides from Adam Lopez [3].

3 Activities

Activities constitute the main differential linchpin of MOOC and its main identity. In addition they represent a breeding ground in order to obtain qualitative and quantitative metrics to further understand the MOOC phenomena.

However, when facing the design of activities, we have to consider that not all MOOC contents are inclusive to all participants. Especially people coming from areas such as linguistics, humanities, fine-arts or translation and interpretation do not have a solid mathematical basis. Thus two non exclusionary are profiles are set to achieve the syllabus certificate of accomplishment. One, i.e. mt-developer, with a complex and strong mathematical background and one i.e., mt-manager, based on a linguistic basis and the understanding of various concepts along with their interaction.

From the first day, teaching assistants will try to identify the roles of leader and follower from within the participants based on different parameters such as participation in the forums, test grading, behavioral pattern both on slide consultation and viewing the video lectures. Leaders will be publicly reputed gaining some extra-motivation with techniques based on authority, popularity and typical metrics from social networks (like, dislike, mention, badges...).

Even so, we set the focus of collective work beyond forum or feedback activities. Therefore, the activities are not only limited to Programming Assignments, Weekly Quizzes but they also try to go deeper on other skills of the student in order to gain joint responsibility of the knowledge path alongside their peers making the learning

http://mt-class.org