A Quick Intensive Course on Natural Language Processing applied to Literary Studies

Borja Navarro-Colorado
Department of Software and Computing Systems
University of Alicante
borja@dlsi.ua.es

Abstract

This paper presents how Natural Language Processing is taught to students of a Master's Degree in Literary Studies. These students' background is solely humanistic, and they have no knowledge whatsoever of Natural Language Processing (NLP). The challenge is to introduce these students to the main aspects of NLP in a 20-hour course, and show them how they can apply these techniques to the analysis of literary texts. The course focuses on three main aspects: first, to get to know a new approach to literary text analysis based on the distant reading model; second, to develop a representative literary corpus and, finally, to apply basic NLP techniques to said corpus in order to extract relevant data. Among these techniques are word frequencies, Part of Speech tagging and distributional semantic models such as LDA Topic Modeling. Satisfaction surveys show that students are satisfied with the course.

1 Introduction

Teaching Natural Language Processing to Literature students is currently a great challenge. Students come to the course with good skills for close-reading literary text analysis and a good background in history of literature and even in literary theory. However, they don't have enough technological or mathematical background in order to understand how current Natural Language Processing techniques work and how they can be applied to the analysis of literary texts. Indeed, at the beginning, they are unsure about the usefulness of these resources for literary studies.

In this paper I will present the objectives and contents of a Master's course (two credits) focused on the application of computational techniques (mainly Natural Language Processing) to literary text analysis. The subject is framed in Moretti's distant reading model (Moretti, 2007; Moretti, 2013; Jockers, 2013) because I think that it is within this approach that NLP techniques are really useful for literary studies. Following a standard empirical text-analysis process, the course is organized in two main modules: the first one is devoted to corpus design, compilation and annotation; and the second one is devoted to the application of some specific Natural Language Processing techniques such as, among others, lexical frequencies analysis, part of speech tagging, named-entities recognition or distributional semantic analysis (LDA Topic Modeling (Blei et al., 2003)). My procedure in class is as follows: first I show my students how each of these above-mentioned techniques work, and then what can be expected from them when applied to literary texts. This way students extract empirical data from the corpus that they must interpret according to their literary knowledge.

Students usually pass the course without much difficulty. Satisfaction surveys show that the course is well received among students. In general they assume the necessity of empirical data to complete traditional literary analysis. However, only a few students eventually apply some of these NLP techniques in their final Master Thesis or PhD Thesis.

In the next section I will present first the course context and the student profile; then I will show the main objectives of the course, how the content is organized and how it is taught to students (theory and practice); to conclude I will propose some ideas for a Digital Humanities curriculum based on this experience.

2 Course context and student profile

The course is called "Computer resources for literary research".[1] It is a twenty-hour course included

[1] The course is taught in Spanish. The exact name is "Recursos informáticos para la investigación literaria": http:

in the Master's Degree on Literary Studies[2] at the University of Alicante (Spain). It is taught face-to-face in a computer lab classroom, where students perform their tasks under the teacher's supervision. The only task that is done outside the computer lab is the students's final essay in which they must apply the NLP techniques they have learned during the course.

The course is taught by a teacher whose background is Spanish Language and Literature (B.A.) with a PhD on Natural Language Processing. So far it has been taught three times since 2014-2015 school year. The first year the course had an attendance of 11 students, 12 students the second year and finally 17 students this last year.

The students who take this course are usually young graduates in Literature. All of them share a good background in humanities and history of literature, and they use similar research methods for the traditional analysis of literary texts. They differ in the literary tradition that they have studied. Most of them are graduates in English Literature or Spanish (Castilian) Literature, but there are also graduates in other literary traditions such as Catalan, Arab or French Literature. Some of them have background in Linguistics as well. In this context, the course is focused mainly on the computational analysis of Spanish (both Catalan and Castilian) and English literary texts.

The students's knowledge about mathematics or computers is poor; as far as mathematics is concerned, their knowledge basically comes down to what they learned in high school. As regarding computers, they are digital natives and use computers in their daily life. However, they have not knowledge at all about Computer Science: algorithms, programming, etc.

On the other hand, these students are familiar not only with the main concepts of Linguistics, but also with the literary criticism models that apply linguistic techniques to literary analysis (such as Russian Formalism, Structuralism or New Criticism). Therefore, they clearly understand the linguistic aspects of Natural Language Processing and its main problem (the linguistic ambiguity). However, their lack of a thorough computational knowledge makes it hard for them to understand how NLP works, that is, the mathematical basis of NLP. During the course not only do I explain how to use NLP tools, but also I try to clarify how they work, that is, how these tools deal with linguistic ambiguity.

3 General objectives

In only 20 teaching hours it is not possible to introduce Python nor any other programming language in the course. This limitation leaves most Natural Language Processing tools out of the syllabus. Moreover, I avoid focusing the course only on the technical application of NLP tools. More than this, students must understand the important contribution of these tools to literary studies, mostly because before they take this course, they do not see why they must apply computational tools to the analysis of literary texts. They have enough with their (manual and close reading) methodological skills and literary analysis models, so they do not see the usefulness of computational analysis. If I want to analyze the metrical aspects of García Lorca's "Little Viennese Waltz", why do I need a computer, when I can analyze properly all these lines by hand? This question is related to the usefulness of NLP for literary studies.

Our first objective is to show that the application of NLP techniques to literary text analysis makes sense only if by using these techniques I can learn something new about the literary phenomenon. The application of NLP tools to emulate human analysis makes no sense. On the contrary, it must be applied where manual analysis cannot reach.

In this regard, Moretti's Distant Reading model (Moretti, 2007; Moretti, 2013; Jockers, 2013) sets up a framework where the computational analysis of literary texts is not only useful but also necessary. I am referring to the computational analysis of large corpora in order to extract common patterns and regularities from the texts and, in general, implicit and unknown information that cannot be extracted by means of a manual analysis. Of course it is better to analyze manually the metrics of García Lorca's "Little Viennese Waltz", but it is not possible for a human being to analyze the whole metrics of all Spanish Golden Age Poetry (all the Spanish poetry composed during the 16th and 17th centuries). In this case the usage of computational analysis and NLP techniques is mandatory, and it will probably show some regularities about the period that traditional approaches are not able to detect. Both approaches are, in the end, complementary.

//www.dlsi.ua.es/~borja/riilua/

[2] https://maesl.ua.es/index.html

The second objective of the course is to formulate big questions. In order to apply the Distant Reading model, students must first learn how to formulate big literary questions, questions that could be answered applying NLP techniques to large literary corpus. At the beginning students pose small questions as the base for the analysis of a single novel or the work of a specific poet. I encourage students to think of big literary questions: Not questions about a specific author or a specific piece of literary work, but questions about whole literary periods or genres, for example.

To develop this new point of view, I use the easy but powerful Google Books n-gram viewer.[3] It allows the student to look for word and n-gram frequencies on the Google Books collection and display them in a timeline. With this tool students practice how to think big. They formulate big questions about literature or, in general, cultural aspects and then look for data in the Google n-gram tool. They then analyze the data provided by the tool and try to answer the question. The kind of questions formulated are based on Michel et al. (2011).

Finally, the third main objective of the course is to show that the application of these techniques sometimes provides quantitative data that, rather than answers, produce new research questions that must be studied (Moretti, 2007).

Once students accept these ideas they are ready to learn about the technical aspects of NLP. Now they are able to appreciate the usefulness of NLP for literary studies and the course makes sense for them.

4 Content and lessons

Content and syllabus are based on my own research experience. This is why the content of the course is structured following a standard empirical text analysis. Given that the main objective of the Master's Degree is to prepare students for research in literary studies, this structure fits well with student expectations. In any case, content and lessons of this course are not based on any specific previous course. Besides my own experience, to set up the course content I have taken into consideration tutorials such as (Manning, 2011), handbooks such as (Jockers, 2014; Pustejovsky and Stubbs, 2013; Jurafsky and Martin, 2008), and some courses on Corpus Linguistics such as (McEnery, 2013) or on Natural Language Processing such as (Jurafsky and Manning, 2012).

The syllabus of the course is as follow:

- Introduction. Objectives (2 hours).
- Module 1. Corpus compilation.
    1. Corpus design and compilation (2 hours).
    2. Corpus annotation (4 hours).
- Module 2. Corpus analysis.
    3. Frequencies, n-grams and concordances (2 hours).
    4. Regular Expressions (2 hours).
    5. Natural Language Processing (4 hours).
    6. Text mining (4 hours).

The main idea of the first module is that only with a representative literary corpus is it possible to achieve reliable conclusions. The literary analysis depends, eventually, on the quality of the corpus and the annotation. In this module students learn about basic aspects of Corpus Linguistics: how to select representative texts according to a set of objective criteria; how to find, download and clean texts in order to obtain plain texts; how to store text files; how to deal with textual codification problems, etc. (Wynne, 2004; Bowker and Pearson, 2002; McEnery and Hardie, 2012)

The second lesson of this module is an introduction to manual corpus annotation. It includes such topics as XML and TEI (TEI Consortium, 2016), the application of annotation guidelines so that the resulting annotation is consistent and reliable, or the evaluation of the annotation through inter-annotators agreement (Pustejovsky and Stubbs, 2013).

Along with this lesson I develop a simulation of a corpus annotation process. As a main resource I use the Corpus of Spanish Golden-Age Sonnets (with metrical annotation) (Navarro-Colorado et al., 2016).[4] This corpus is suitable for this exercise because it is freely available (including the annotators's guidelines), it follows the standard XML-TEI, and it has been manually annotated with literary information: the metrics of each line. This corpus provides ample annotation practice, and eventually the students can compare their own work with the original corpus annotation.

[3] https://books.google.com/ngrams

[4] https://github.com/bncolorado/CorpusSonetosSigloDeOro

The second module is focused on the computational analysis of a literary corpus. It is structured in four lessons.

The objective of the first lesson (number 3) is to set up the basis for the computational treatment of texts. Specifically, I show students how words are transformed into numbers and what these numbers represent. With AntConc tool (Anthony, 2014),[5] students perform several tasks such as: the extraction of the most frequent tokens of the corpus (including a stop-words filter), the extraction of the most frequent n-grams, the estimation of the type/token ratio, concordance analysis, or the extraction of the most frequent lemmas. The main conclusion of this lesson is that, when the corpus is really large, it is difficult to extract generalizations from it using these techniques (Roe, 2012).

Lesson 4 is devoted to a gentle introduction to regular expressions. The objective is to show students how to formalize linguistic expressions. Although it is not possible to go deeply into this topics, students learn how to define regular expressions that allow them to find words by stem or by rhyme, or even conditional expressions (words that appear before or after another word, etc.).

The lesson devoted to Natural Language Processing techniques (lesson 5) is focused on part of speech (PoS) taggers, syntactic parsing and named-entities (NE) recognition. In general, I show first the main architecture of this kind of tools, then the main problems (ambiguity) and finally the common error rate.

In the case of the part of speech tagger, for example, I explain that each word-lemma is related to all its possible parts of speech in a dictionary. This way the PoS ambiguity problem is presented. Then some standard solutions are explained, as the use of a set of rules to specify the suitable part of speech for each word in each context, or the application of statistical information. Named-entities recognition is explained in the same way. Syntactic parsing is explained showing how Context Free Grammars (CFG) and Probabilistic CFP work.

In any case it is not our objective to explain deeply the solutions to these problems (grammar development, statistical learning, machine learning, etc.). These concepts will be hard to follow for our students. What is most important is that students understand the computational problem. This way they will apply these techniques knowing what they can expect from them.

The exercises in this lesson are carried out with FreeLing (Padró and Stanilovsky, 2012). This tool is appropriate for our course because it is multilingual: It includes PoS tagger, NE recognition and chunkers for Spanish, English, Catalan and other languages. The drawback of FreeLing is that it is hard to install and use. Among other things, it has no graphical interface. To avoid installation problems, instead of using FreeLing directly we use a web application developed by Pompeu Fabra University (UPF - Barcelona, Spain) called ContaWords.[6] Using this web application is very easy: It allows students to upload several text files that are analyzed by FreeLing on a remote server. Results are returned in a spreadsheet format. To obtain the data in a spreadsheet is a must: Students can create graphics and analyze the data extracted directly from their corpus.

The final section of this module is devoted to computational semantics (lesson 6). Among all the different computational semantic models that have been proposed (lexical semantics, first-order logic, events and semantic roles, etc.), our course is solely focused on distributional models of semantics. I explain only this model because it allows an efficient computational processing of large corpora, and because it can only be used with computers.

Two main theoretical concepts are explained in this lesson: first the idea lying behind distributional semantics that the meaning of a word depends on its contexts and words that occur in similar contexts have similar meaning (Harris, 1968), and second how computers deal with word contexts by means of vectors and matrices (Turney and Pantel, 2010). How it is possible to know the semantic similarity between two words in the distributional framework is shown following Widdows (2004).

In order to show students how distributional models of semantics (and, in general, text mining techniques) are able to extract generalizations and regularities from large corpora, I explain LDA Topic Modeling (Blei et al., 2003). I describe how it works and how it can cluster words with similar (distributional) meaning in the same topic.

Once students understand how LDA works, it is applied to a literary corpus using MALLET (McCallum, 2002). Students must extract a set of topic

[5] http://www.laurenceanthony.net/software/antconc/

[6] http://contawords.iula.upf.edu/executions

models and analyze them. They check if topic models are coherent, and if the words grouping in each topic could be justified by means of literary criteria. As I said before, I encourage students to formulate questions (as, for example "why words A and B are in the same topic?") and try to answer them according to their background in literature.

5 Evaluation

With the exception of MALLET, which is the most complex tool used in the course, students do not have much difficulty using the technology and completing the exercises in the syllabus, and eventually they all pass the course.

In order to monitor courses, the University of Alicante distributes a survey in order to know the degree of satisfaction of the students with the course taken. This course was marked with 9 points out of 10, showing that students are really satisfied with the course. For us, these data show that the approach used to teaching NLP in literary studies is appropriate. However, only a few students apply some of these techniques in their final Master Thesis or PhD Thesis. Perhaps students need more time to assimilate all these new techniques and apply them to their daily research in literature.

6 Conclusions

In this paper I have presented the key points of our approach to teaching NLP to students of literature. I try to open the students's mind with these main ideas:

1. The use of NLP techniques in literary studies makes sense when they are applied where manual analysis cannot reach (the analysis of large literary corpora).

2. This literary analysis approach requires a wide scope -students must extend their point of view and learn how to formulate big research questions.

3. The application of these techniques sometimes provides data that, rather than giving answers produces new research questions.

The course is structured following a standard empirical text analysis: compiling, first, a representative literary corpus, and then analyzing it with NLP techniques (frequencies, part os speech tagging, LDA Topic Modeling, etc.). As the course has only 20 hours, I do not go deeply into the technical details of NLP. I first explain how each technique works and then students apply them to the corpus with easy-to-use tools.

Acknowledgments

I would like to thank the anonymous reviewers for their helpful suggestions and comments. Thanks also to my students for their feedback that helps me to improve the course.

Paper partially supported by the BBVA Foundation: grants for research groups 2016, project "Distant Reading Approach to Golden-Age Spanish Sonnets"

References

Laurence Anthony. 2014. Antconc (version 3.4.3) [computer software]. http://www.laurenceanthony.net/. Tokyo, Japan: Waseda University.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Lynne Bowker and Jennifer Pearson. 2002. *Working with Specialized Language. A practical guide to using corpora.* Routledge, London.

Zellig Harris. 1968. *Mathematical structures of language*. Wiley, New York.

Matthew L. Jockers. 2013. *Macroanalysis. Digital Media and Literary History.* University of Illinois Press, Illinois.

Matthew L. Jockers. 2014. *Text Analysis with R for Students of Literature*. Springer, Switzerland.

Dan Jurafsky and Christopher Manning. 2012. Natural language processing. http://online.stanford.edu/course/natural-language-processing. Stanford University.

Dan Jurafsky and James H. Martin. 2008. *Speech and Language Processing*. Prentice Hall.

Christopher Manning. 2011. Natural language tools for the digital humanities. https://nlp.stanford.edu/ manning/courses/DigitalHumanities/. Stanford University.

Andrew K. McCallum. 2002. Mallet: A machine learning for language toolkit.

Tony McEnery and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theorie and Practice*. Cambridge University Press, Cambridge.

Tony McEnery. 2013. Corpus linguistics: Method, analysis, interpretation. https://www.futurelearn.com/courses/corpus-linguistics. Lancaster University.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(176).

Franco Moretti. 2007. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.

Franco Moretti. 2013. *Distant reading*. Verso.

Borja Navarro-Colorado, Mara Ribes Lafoz, and Noelia Snchez. 2016. Metrical annotation of a large corpus of Spanish sonnets: representation, scansion and evaluation. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, Slovenia.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Language Resources and Evaluation Conference (LREC 2012)*, Istanbul.

James Pustejovsky and Amber Stubbs. 2013. *Natural Language Annotation for Machine Learning*. O'Reilly.

Glenn Roe. 2012. The dangers and delights of data mining. In *Digital Humanities Summer School*, Oxford (UK). University of Oxford.

TEI Consortium, editor. 2016. *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.1.0. Last modi ed 15th December 2016*.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Dominic Widdows. 2004. *Geometry and Meaningn*. CSLI publications.

Martin Wynne. 2004. Developing Linguistic Corpora: a Guide to Good Practice. http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm.