Ambiruptor: The Ambiguation Interruptor

Project Proposal

Boumediene Brikci Sid
Ievgeniia Oshurko
Maria Boritchev
Pierre Ohlmann
Samir Tendjaoui
Simon Mauras
Thi Xuan Vu
Victor Hublitz

22th October 2015

Abstract

Throughout the second year at the ENS Lyon (first year of master), students have to work on an integrated project. This paper describes the proposal of the Ambiruptor project which is born from the merge of the NER project and the Joke Generator project.

The main objective is to produce an efficient tool able to give the correct meaning of an ambiguous text. Several machine learning concepts are going to be implemented, and trained on annotated texts from the Wikimedia database.

Contents

1 Key parts

1.1 Research

The first part of our work is going to be research. You can find in section 3 a first list of articles. During the research part we are going to split in 3 groups working on the following topics:

- Usual technics for text disambiguation
- Machine learning algorithms.
- Existing tools for data mining.

The first topic is mainly "State of the art" research. The goal is to find what has already been done and what is currently studied. The purpose of the second group is to compile a list of machine learning technics that can be used to solve our problem. The third group is going to look for tools that can enable us to mine workable data from the internet (Python API ?).

At the end of this part, each group is going to do a quick summary on their results, so that everyone have an overview on the project.

1.2 Design

During the second part we are going to think about the structure and the design of the tool we want to implement. To do so we will split in 2 groups considering several possible ways on the following topics:

- How should we implement the Ambiruptor tool ? (Languages, modules, ...)
- How should the data be represented ? (Data storage, data-structures, ...)

Those two groups will exchange informations to make sure that their choices are compatible. The deadline for the end of this part is fixed on **November the 26th, 2015**.

1.3 Intermediary report

The intermediary report's deadline is fixed on **December the 3rd, 2015**. It will clarify all the details of the upcoming implementation. A small group of two or three people will summarize the decisions taken during the design part.

1.4 Implementation

The third part is probably the longest one. The implementation will be divided in 2 groups:

- Adpatation of the chosen tools to extract workable data from the internet.
- Implementation of the different modules (defined during the design part)

The progress of the second group slightly depends on the results of the first group. The deadline for the end of the implementation part is fixed on **January the 20th, 2016**.

1.5 Testing

Our tool will be tested progressively during the implementation. Yet, this part only consists of an industrial scale test phase. We will launch our tool on all the data we are able to process and it might take a while. The efficiency of our implementation will be calculated.

1.6 Addons

As soon as the Ambiruptor tool becomes functional, the creation of user-friendly interfaces will be assign to a small group of students.

- Web application
- Pdf reader plugin

We can expand this part if the project goes well and fast. Likewise, this part can be shortened if we get stuck in one of the previous parts.

The deadline for implementation, testing and addons is fixed on **March the 23th, 2016**.

1.7 Final report

The final report's deadline is fixed on **April the 15th, 2016** (one month before the public presentation). All students will summarize their year's work. Two or three people will compile those documents and produce the final report.

2 Gantt diagram

cf diagram Victor

3 References

cf articles sent on the mailing list