Ambiruptor: The Lexical Ambiguation Interruptor

Project Proposal

October 21, 2015

Presentation

The main objective is to produce an efficient tool that gives the correct meaning of an ambiguous text. Our tool will be based on annotated texts from Wikimedia (WikiMedia or WikiPedia?) database. We are going to implement several machine learning concepts to train our tool and to solve the ambiguities in texts.

Our team contains 8 members: Maria Boritchev, Boumediene Brikci Sid, Victor Hublitz, Simon Mauras, Pierre Ohlmann, Ievgeniia Oshurko, Samir Tendjaoui, Thi Xuan Vu. The coordinator (team-leader/big boss/any other cool word) of the project is Simon Mauras (??) definitely not me :).

Contents

Motivation

Please, put smth here. Like what is the disambiguation, maybe a small example (the one with bar?). And the most important: who cares about it besides of you 8 :)

1 Objectives

- Data retrieving/mining from existing sources: Wiki(p|m)edia, if some time left Thesaurus/WordNet
- Data processing/analysis, like put some labels on the words, or cluster the words/senses/parts of speech/whatever. Probably you'll need to create your own model (or format) of data, might be json or xml.
- Create your own training base/your dictionary. Should be easily modifiable
- Learn some disambiguation methods and machine learning algorithms. Develop the core for your ambiguity resolving tool. If you choose this, or if time left using the supervised machine learning would be great. The last is completely Olga's opinion, if you are disagree, we forget about it :)
- Implement some testing module
- When everything is ready, put it on the Web, create a website, etc.

You might also create several lists, like main objectives and secondary objectives (in case if you have too much time).

2 Plan of Our Work

2.1 Research on the State of the Art

The first part of our work is going to be research. You can find in Section 5 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 text above does not sound very sexy, try to reformulate. I'm not sure if you need a list of articles here

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 several 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.

2.2 Design

During the second part we are going to think about the detailed structure and the design of the software we want to implement. Olga will put a preliminary picture as soon as gets home. If you don't like it, feel free to change it. I'm not a dictator, just want to propose you smth

The deadline for the end of this part is fixed on **November the 26th, 2015**.

2.3 Intermediate report

We set the deadline for our intermediate report to **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.

2.4 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). I don't like this name, too generic

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**.

2.5 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 tell how!.

2.6 Addons

As soon as the Ambiruptor tool becomes functional, the creation of user-friendly interfaces will be assigned 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**.

2.7 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.

Comment on the whole section: Do you need the sentences like "the deadline is fixed on bla". Just put your Gantt chart, everybody can read it. There is no need to make the paper too long by using such sentences, be short and efficient

3 Gantt chart

cf diagram Victor

4 What about Milestones?

5 References

cf articles sent on the mailing list