Ambiruptor
The Lexical Ambiguation Interruptor

Final Report

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

December 22, 2015

Abstract

The main point of our project is to develop a word-sense disambiguation tool. Our aim is to be able, given a certain text, to determine the actual meaning of each ambiguous word. To this end we use Wikipedia, and more specifically its internal links, in order to produce an annotated corpus from which a machine learning framework is developed. At the present time, we are clear about our objectives and we have chosen the abstract design of our future code. Furthermore, the coding part is now well advanced as we achieve mining Wikipedia and have already implemented several feature extractors.

Contents

1   Presentation

Word Sense Disambiguation is a Natural Language Processing task that lies in the assignment of the appropriate meaning to a word according to a given context, and its separation from other possible meanings. Since the 1940s, this problem has proved its difficulty and the lack of database has forced people to label each word manually.

Nowadays, the Internet creates new possibilities to get sufficiently big databases and the use of new machine-learning methods has given more efficient results on this open problem.
There are several possible applications of Word Sense Disambiguation:

- Machine Translation

- Information Retrieval

- Semantic Parsing

- Speech Synthesis and Recognition

1.1   Ambiruptor project

The main objective of the **Ambiruptor project** is to produce an efficient tool that gives the correct meaning of ambiguous words in a text. Our tool will be based on several supervised machine learning concepts. We use Wikipedia to build our learning corpus and to annotate it according to its internal links.
All the code we produce is under the GNU GPLv3 license.

1.2   Work team

Our team consists of 8 master students of the ENS of Lyon: Boumediene Brikci Sid, Maria Boritchev, Victor Hublitz, Simon Mauras, Pierre Ohlmann, Ievgeniia Oshurko, Samir Tendjaoui and Thi Xuan Vu. The coordinators of the project are Simon Mauras and Ievgeniia Oshurko.