String Edit Distance Module (sed)

Execution reference

Jose Oncina
Cristian Olivares-Rodriguez

June 21, 2007

1 Directory structure

In each directory they are the source codes of the components of this module and they are organized of the following way:

Synthetic experiments components:

- **/automaton** : this component contain the source code of the random automaton that generates random strings.
- **/strings** : this component contain the source code that generates random strings from some vocabulary size and string length.
- **/rtransducer** : this component contain the source code that generates a random transducer
- **/db** : this component contain the source code that builds a string pairs database from generated strings in the first component and the random transducer.
- **/distance** : this component contain the source code that compute the distance between both conditional transducer and joint transducer to the random transducer.

Real experiments components (Handwritten digits):

- **/divide** : this component contain the source code that allows to divide the digits database into both learning and test set.
- **/pairs** : this component contain the source code that obtain the nearest neighbor to each sample of training set.
- **/test** : this component contain the source code that compute the goodness of the developed transducers in relation with the classic edit distance.
- **/digits** : this directory contain digits database that have been used in the real experiments.

Shared components:

- **/sed** : this component contain the source code that allows to generate so much a conditional transducer as one joint transducer, taking as input the database generated before.

Others directories:

- **/src** : this directory contain the source codes of the shared artefacts by all the components.
- **/tmp** : this directory stores the generated temporary files in the process.
- **/bin** : this directory contain the executable files.
- **/lib** : this directory stores the shared library.

2 Execution

The module executes a synthetic experiment on both conditional and joint transducers from generated data by a random transducer, that is to say, synthetic data.

- In order to execute the predefined synthetic experiments it is necesary:
  1. To decompress the file *Stochastic-0.0.9.tar.gz*.
  2. To execute the shell script *synthetic.sh*, located in the **/bin** directory, as it follows: **$ synthetic.sh -v** [**vocabulary-size**]
  3. To analyze the distances files located in the **/tmp** directory. These files store the results of each computed distance.
- In order to execute the predefined real experiments it is necesary:
  1. To decompress the file *Stochastic-0.0.9.tar.gz*.
  2. To execute the shell script *character_hw.sh*, located in the **/bin** directory, as it follows: **$ character_hw.sh -a**
  3. To analyze the results files located in the **/tmp** directory. These files store the results of each computed distance.

*** It's important to say that the runtime of this experiments is very large.**

- In order to execute the only one synthetic experiment it is necesary:
  1. To decompress the file *Stochastic-0.0.9.tar.gz*.
  2. To enter the **/bin** directory.
  3. To generate the random strings with the *automata* component.
  4. To create the random transducer with the *rtrans* component.
  5. To generate the string pairs database with the *db* component.
  6. To create the conditional transducer with the *sed* component and using the anyone *prob* file located in the **/tmp** directory.
  7. To create the joint transducer with the *sed* component, the *-na* option equal to *j* and using the anyone *prob* file located in the **/tmp** directory.

8. To compute the distance between random transducer and conditional transducer with the *distance* component.
9. To compute the distance between random transducer and joint transducer with the *distance* component and the *-d* option equal to *j*.

- In order to execute the only one real experiment it is necesary:

1. To decompress the file *Stochastic-0.0.9.tar.gz*.
2. To enter the **/bin** directory.
3. To divide the digits database with the *divide* component.
4. To create the pairs of strings with the *pairs* component.
5. To create the conditional transducer with the *sed* component and using the anyone *prob* file located in the **/tmp** directory.
6. To create the joint transducer with the *sed* component, the *-na* option equal to *j* and using the anyone *prob* file located in the **/tmp** directory.
7. To compute the goodness of joint and conditional transducer in relation to matrix of costs that correspond to classic edit distance with the test component.