A Description of Data Format and Usages of the Parser

DATA FORMAT

The data format we use follows the data format of SemEval 2015 Task 18.
A detailed description could be found on this page:
http://alt.qcri.org/semeval2015/task18/index.php?id=data-and-tools
Some trial data for this task could also be downloaded from the above website while the complete data is now released at LDC.

Note that the data reader of my project can read in data only in this format, but can use any corpus if you could transform the format to the SDP style.

Same samples are provided under the directory

/DataSample/

For data without semantic dependency annotation that you want to parse with a saved model, you must preprocess the data to acquire the lemmas and POS tags. Each line of the token should be formatted as below:

id token(form of word) lemma POStag - -

Examples:

With annotation:

```
#sentence_id : sample 1
1  Billing  billing  NN  -  -  ARG2  ARG1  _  _  _  _
2  is  is  VBZ  -  -  _  _  _  _  _  _
3  included  include  VBN  +  +  _  _  _  _  _  _
4  in  in  IN  -  +  ARG3  _  _  _  _  _
5  a  a  DT  -  +  _  _  _  _  _  _
6  caller  caller  NN  -  +  _  _  BV  _  _  _
7  's  's  POS  -  -  _  _  _  _  _  _
8  regular  regular  JJ  -  +  _  _  _  _  _  _
9  phone  phone  NN  -  +  _  _  _  _  _  _
10  bill  bill  NN  -  -  _  ARG2  _  poss  ARG1  compound
11  .  _  .  -  -  _  _  _  _  _  _
```

No annotation:

```
#sentence_id : sample 1
1	Billing	billing	NN	-	-
2	is	is	VBZ	-	-
3	included	include	VBN	-	-
4	in	in	IN	-	-
5	a	a	DT	-	-
6	caller	caller	NN	-	-
7	's	's	POS	-	-
8	regular	regular	JJ	-	-
9	phone	phone	NN	-	-
10	bill	bill	NN	-	-
11	.	_	.	-	-
```

USAGE

Users should give file paths before training a model or using a saved model.

1) Set file paths and parameters in /TransitionBasedParser/config.h

Modify contents in the section between line 18 to line 44:

```
18 /****************************************** SETTINGS ******************************************/
19 /****** Set file paths and global constants before compiling and running the project. ******/
```

/* MODE */

Here, you should decide whether to train a new model or use a saved model.
If you are training a new model, you should also decide whether use label and whether consider the virtual arc from ROOT of dependency graphs.

/* FILE PATHS */

Appoint file paths here. A complete model is consist of:
5 dictionaries:
map words, lemmas, POS tags, transitions, features to *int* respectively. The paths for those dictionaries are: WORD_DICT_PATH, LEMMA_DICT_PATH, POS_DICT_PATH, TRANSITION_DICT_PATH, and FEATURE_DICT_PATH
and 1 file to save parameters of the perceptron.

When training a new model, you should appoint the above 5 dictionary paths and CHECKP_PATH in which directory parameters of perceptron are saved after finishing each epoch during training.

TRAIN_DATA, TEST_DATA, and RESULT_FILE, are the paths of training data, test data, and the parsing result parsed with the obtained model when finishing training.

When using a saved model, you should appoint the above 5 dictionary paths and SAVED_ALPHA which is the file containing saved parameters of perceptron. TEST_DATA, RESULT_FILE are the paths of test data, and the file to write result.

/* PARAMETER SETTING */

Set the width of beam. And decide how many epochs are used if train a model.

2) Then you can compile and run the project.