Karlsruhe Institute of Technology
Dr. Eunah Cho
Institute for Anthropomatics and Robotics, Interactive Systems Labs
eunah.cho@kit.edu

MT Praktikum - Word Embeddings & NNLM

4. Juli 2017

Um die Aufgaben auszuführen, können Sie Ihre Daten in folgendem Verzeichnis speichern:
`/project/smtstud/ss17/systems/USERNAME/`

We are going to use a pre-trained word vectors. Copy this word vector file into your directory.

```
/project/smtstud/ss17/data/vec/Wemb.en.filtered.lowered
```

Later you are going to calculate the cosine similarity of words represented in vector space. Copy the script into your directory.

```
/project/smtstud/ss17/bin/vector.py
```

A. Word Embeddings

1. In the last page, you can find visualization from word embeddings obtained from English and French machine translation data. It is also available at http://i13pc106.ira.uka.de/∼echo/research.pdf. Words are filtered under the topic “Research”.

   - Find at least three groups where you can see the similarity in meanings and mark on the visualization. You can also check a English-French dictionary: http://enfr.dict.cc/.

2. We are going to examine pre-trained 200-dimensional vectors for 10K English vocabulary. Examine the format of the file.

   Word vectors represent semantic and syntactic relationship between words. For example, *lady* should be closely related with *woman.* Check their vector values.

   - Check vector values of the word *lady* and *woman.*

word	v_0	v_1	v_2	v_3	...	v_{199}
lady				...		
woman				...		

3. We can check the similarity of the words by loading the vectors and calculating the distance between them.

```
python vector.py
```

- What is the most similar word to the word *lady*? What is their similarity?
- Find the most similar word for following words and fill the table.

Input word	Most similar word	Similarity
lady		
book		
mother		
school		
great		
tree		

- Find the similarity score of following words and fill the table. When is the similarity score high?

word A	word B	similarity
father	dad	
human	animal	
apple	big	
soccer	football	

- Find the word which fits in the semantic relationship.

man : husband = woman : wife
grass : green = sky :
tree : forest = water :

B. Neural Language Model

1. Take a look into the log file of training a neural language model.
   `/project/smtstud/ss17/data/rnnlm/Train.log`

   - How is the perplexity score for development data?

2. There are four RNN language models trained. You can find them in
   `/project/smtstud/ss17/models/rnnlms/`

   All models are trained with top 5,000 words.

   (a) Forward LM, two layers

   (b) Backward LM, two layers

   (c) Forward LM, two layers but half of the size

   (d) Forward LM, one layer

   We are going to apply each LM on the test data.

   `/project/smtstud/ss17/data/rnnlm/test.de`

   - Which sentences in the test data should have lower perplexity? Why?

3. Copy the following script into your directory and run it in your directory.

   `/project/smtstud/ss17/bin/rnnlm/Test.back_forward.sh`

   The script calculates perplexity for each sentence using the backward and the forward LM.

   - How is the perplexity for each sentence using each model?

	Forward	Backward
ich melde mich für die Konferenz an .		
ich melde mich für die Konferenz auf .		
ich schlage mit der rechten Hand auf .		
ich schlage mit der rechten Hand vor .		
mein Freund , den ich seit vielen Jahren kenne , ist nach Stuttgart gezogen .		
mein Freund , den ich seit vielen Jahren kenne , sind nach Stuttgart gezogen .		
die Verkäuferin ist nett .		
die Marklerin ist nett .		

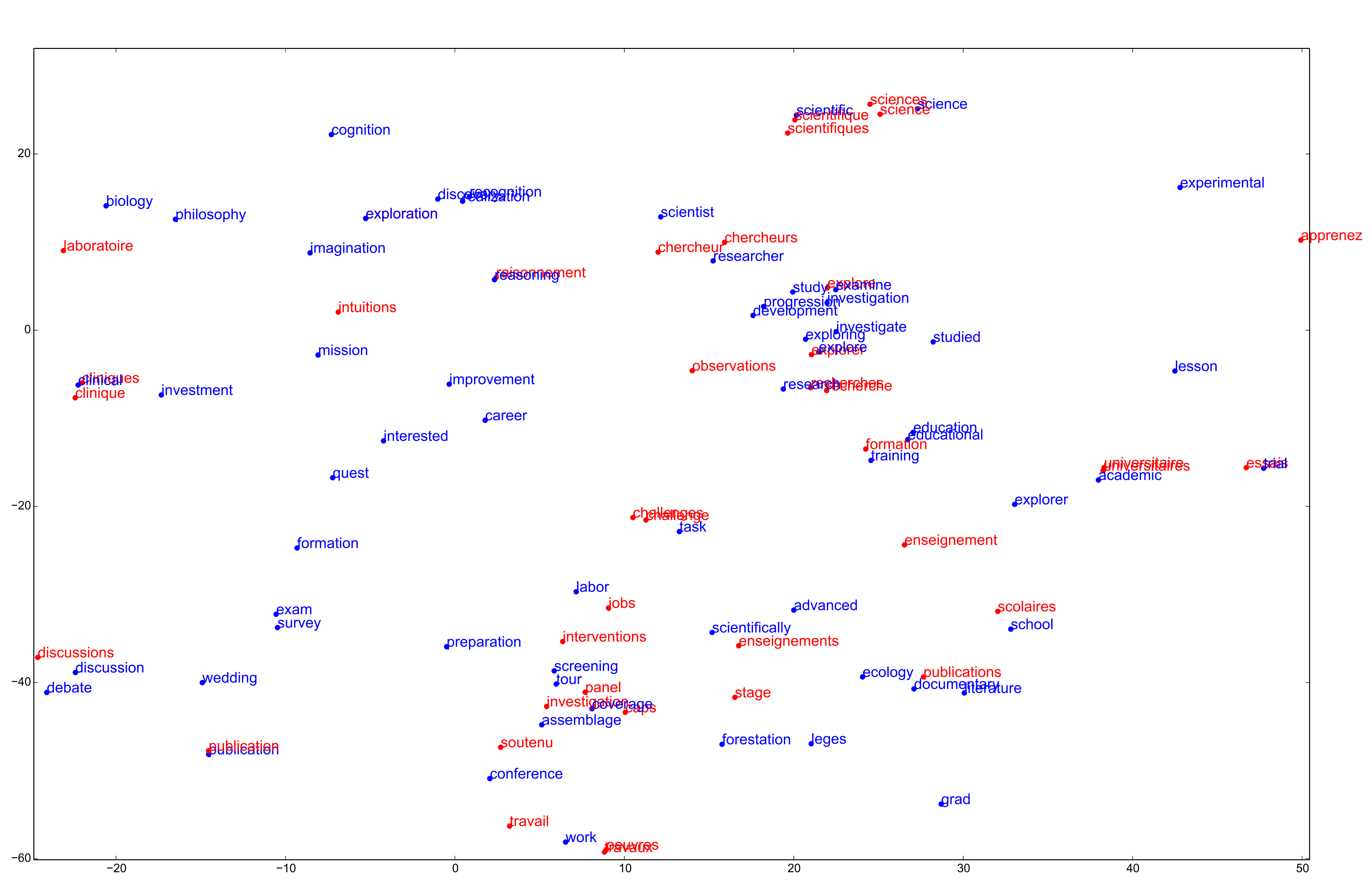
4. For the same test data, try an LM that has half-sized dimensions and another LM that has only one layer. Copy the following script into your directory and run it in your directory.

   `/project/smtstud/ss17/bin/rnnlm/Test.halfdim_onelayer.sh`

   - How is the perplexity for each sentence using each model?

	HalfDim	OneLayer
ich melde mich für die Konferenz an .		
ich melde mich für die Konferenz auf .		
ich schlage mit der rechten Hand auf .		
ich schlage mit der rechten Hand vor .		
mein Freund , den ich seit vielen Jahren kenne , ist nach Stuttgart gezogen .		
mein Freund , den ich seit vielen Jahren kenne , sind nach Stuttgart gezogen .		
die Verkäuferin ist nett .		
die Marklerin ist nett .		

5. Feel free to try your own examples by inputting your own `testdata` in the bash file.


sciences
science
scientific
scientifique
scientifiques
cognition
experimental
biology
philosophy
discovery
recognition
realization
exploration
scientist
chercheurs
chercheur
researcher
apprenez
laboratoire
imagination
raisonnement
reasoning
study
explore
examine
investigation
progression
development
intuitions
investigate
exploring
explore
explorer
studied
mission
observations
lesson
cliniques
clinical
clinique
investment
improvement
research
recherches
recherche
career
interested
education
educational
formation
training
universitaire
universitaires
academic
essai
trial
quest
explorer
challenges
challenge
task
formation
enseignement
labor
jobs
exam
survey
advanced
scolaires
school
scientifically
preparation
interventions
enseignements
discussions
discussion
wedding
screening
tour
panel
ecology
publications
documentary
literature
debate
investigation
coverage
caps
stage
assemblage
publication
publication
soutenu
forestation
leges
conference
grad
travail
work
oeuvres
travaux
20
0
-20
-40
-60
-20
-10
0
10
20
30
40
50