Lab 2 – Word-based SMT

Sara Stymne

2017-04-12

Lab 2

- Aim:
  - Understand word-based SMT
  - Gain an understanding of the role of the translation and language models
- Part 1 (today)
  - Play around with probabilities
  - Evaluate the effects
  - Discuss the results in class
- Part 2 (April 24)
  - Train a proper LM
  - Evaluate the results
  - Write a report (April 28)

Domain and Languages

- Block world domain
  - take an arrow
  - put the red block on the circle
  - she puts her red block on the circle
- Focus on 10 given test sentences

Domain and Languages

- Block world domain
  - take an arrow
  - put the red block on the circle
  - she puts her red block on the circle
- Focus on 10 given test sentences
- Swedish ↔ English
  - Focus on analyzing translations into your strongest language!
  - Swedish grammar sketch available

Language Issues

- Verb inflection
  - En: subject-verb agreement
  - Sw: imperative vs present tense
- Noun phrases
  - En: a/an
  - Sw: agreement between articles, adjectives, and nouns
  - Sw: definiteness
  - Different systems for possessives

Decoder and Models

- Simple word-based decoder without reordering
- Translation models:
    - Word translation model
    - Fertility model
- Language model

Decoder and Models

- Simple word-based decoder without reordering
- Translation models:
    - Word translation model
    - Fertility model

- Language model
- The models are given with uniform probabilities
- All necessary words and word translations are given

Word Translation Model

- Contains probabilities for the translation between two words
- tmw.engswe tmw.sweeng

```
the    den      0.33
the    det      0.33
the    NULL     0.33
blue   blå      0.33
blue   blåa     0.33
blue   blått    0.33
block  block    0.5
block  blocket  0.5
```

Fertility Model

- Contains probabilities for the number of words that a word translates into
- tmf.eng tmf.swe

```
take   1 1
takes  1 1
the    0 0.5
the    1 0.5

blocket 1 0.5
blocket 2 0.5
cirkel  1 1
cirkeln 1 0.5
cirkeln 2 0.5
```

Language Model

- Contains probabilities for words and word sequences in the target language
- lm.eng lm.swe

```
\1-grams:
0.04 a
0.04 arrow

\2-grams:
0.04 a block
0.04 an block
0.04 he puts
0.04 he put
0.04 puts he

\3-grams:
add some if you wish!
```

Evaluation

- Since the domain is very simple, and the decoder can output all correct sentences
  - No need to use standard evaluation metrics
  - Evaluate by finding the rank of the correctly translated sentence in an n-best list
- Your tasks
  - Evaluate translation by rank
  - Qualitatively investigate the quality and issues of translation by manually inspecting the translations in the n-best lists

Running the decoder

- Run the decoder and show the N best translations for each sentence
  ```
  /local/kurs/mt/lab2/simple_decoder/translate -lm
  lm.eng -tmw tmw.sweeng -tmf tmf.swe -o 2 -in
  test_meningar.swe -n 25
  ```
- Run the decoder and give the rank of each sentence
  ```
  /local/kurs/mt/lab2/simple_decoder/translate -lm
  lm.eng -tmw tmw.sweeng -tmf tmf.swe -o 2 -in
  test_meningar.swe -eval test_meningar.eng
  ```

What to do today

1. Copy all necessary files
2. Run the decoder in both directions with the given files and analyze
3. Translation model experiments
   - Change some probabilities
   - Investigate the effect
4. Language model experiments
   - Change probabilities
   - Add more 2-grams and/or 3-grams
5. Discussion in full group
   - Take notes during your work!

Time plan

- 9.15 – 11.15
  - Introduction + own work
- 11.15
  - Run blind test (command will be given)
  - Share your scores (in at least one translation direction)
  - Each group discusses some interesting findings
  - Common discussion

Time plan

- 9.15 – 11.15
  - Introduction + own work
- 11.15
  - Run blind test (command will be given)
  - Share your scores (in at least one translation direction)
  - Each group discusses some interesting findings
  - Common discussion
- Ask questions if you get stuck / want to discuss something!
- I will walk around and talk to all groups

Time plan

- 9.15 – 11.15
    - Introduction + own work
- 11.15
    - Run blind test (command will be given)
    - Share your scores (in at least one translation direction)
    - Each group discusses some interesting findings
    - Common discussion
- Ask questions if you get stuck / want to discuss something!
- I will walk around and talk to all groups
- You may take a break when it suits you, but everyone needs to be here at 11.15!

Some Pointers

- Read the instructions, but don't be afraid to ask questions as well!
- Try to think about linguistically motivated changes
  - But you can also try 'crazy' things, just to note the effect
- It is very easy to overfit to the sentences given. Try to do general changes!
  - You will get to evaluate your model on a secret test set at the end of the session
- There are many given questions to think about. You do not have to be able to answer them all, but should be prepared to discuss them at the end of the session
- This lab is only an exercise, normally you would never change probabilities like this!