SMT lectures and labs

Content of the SMT lectures and labs, references

Lecture 4: SMT introduction (Bruno Pouliquen)

- Part 1: Statistical machine translation, different approaches:
  - Syntax-based (also called tree-based)
  - (deprecated: example based)
  - Phrase-based
  - Neural networks
- Which method to choose?
- Various corpora of parallel texts (JRC Acquis, Europarl, UN corpus, Indic corpora…)
- Part 2: Phrase-based SMT
  - History
  - How does it work?
  - First publication Koehn/Och/Marcu (2002)
  - Translation model
  - Language model
  - Different open-sources: Moses / Joshua / Cdec / …
- Part 3: Learning lexical translations
  - IBM model 1
  - Preparation for the lab session
- Part 4: Improved SMT
  - How to improve the "basic" PBSMT
  - Tokenization
  - Normalization
  - Byte pair encoding: reduce vocabulary space
  - Factors in Moses
  - Tree-based (syntax-based) models

References

- ✓ Syntax based MT : http://homepages.inf.ed.ac.uk/pkoehn/publications/esslli-slides-day5.pdf
- ✓ List of MT open-source tools: http://fosmt.org/
- ✓ http://ufal.mff.cuni.cz/mtm16/files/14-phrase-based-smt-ulrich-germann.pdf
- ✓ *[Koehn et al 2002] Statistical Phrase-Based Translation (2002) Philipp Koehn , Franz Josef Och , Daniel Marcu*
- ✓ Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. (2007) "Moses: Open Source Toolkit for Statistical Machine Translation". *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007*

Lab 4: Implement an IBM model 1 model (Bruno Pouliquen)

IBM model 1: A nice example of machine learning: from bilingual texts, learn automatically word translations (using expectation maximization algorithm).
In python/Java/Perl (or other language) develop a program to output lexical probabilities out of parallel sentences

See lecture:

http://www.statmt.org/book/slides/04-word-based-models.pdf (slides 12-30)

Lecture 5: Components of the PBSMT process (Bruno Pouliquen)

- Part 1:Various steps in PBSMT
  - Noisy channel model
  - Model estimation
  - Word translation probability
  - Word alignment probability
  - IBM model 1
  - IBM models 1 to 5
  - Alignment grow-diag final-and
  - Scoring phrase table
  - Distortion model

References

✓ References: original slides by Ulrich Germann:
http://ufal.mff.cuni.cz/mtm16/files/14-phrase-based-smt-ulrich-germann.pdf (p 1-60)

- Part 2: Language models (LM)
  - Language models: How likely is a string of English words good English?
  - N-gram models (Markov assumption)
  - Perplexity
  - Count smoothing
  - Interpolation and backoff
  - Introduction to Neural LM

References:

  - Introduction to language models:
    http://mt-class.org/jhu/slides/lecture-lm.pdf
  - Neural network language models
    http://mt-class.org/jhu/slides/lecture-nn-lm.pdf
  - A complete tutorial from Kenneth Heathfield
    http://ufal.mff.cuni.cz/mtm16/files/08-n-gram-language-modeling-including-feed-forward-kenneth-heafield.pdf
  - A tutorial about LNN and word embeddings
    http://ufal.mff.cuni.cz/mtm16/files/09-nn-language-models-david-vilar.pdf

- Part 3: Decoding algorithm
  - Overview of the decoding (translation)
  - Hypothesis expansion / recombination
  - Beam search
  - Future cost estimation
  - Word lattices & n-best lists

Reference:

  - Tutorial by Ulrich Germann, Philipp Koehn & Mattias Huck
    http://ufal.mff.cuni.cz/mtm16/files/14-phrase-based-smt-ulrich-germann.pdf (p. 61-)

- Part 4: MT Preparation steps
  - Preparation steps (cleaning, sentence aligners, tokenizer)
  - Post-processing steps (e.g. recaser)
  - Introduction to CAT tools

Lab 5: Creating a real SMT model
  - Build a "toy" model out of existing bitexts (sentence aligned corpus, tokenized) (e.g. English-Hindi) on a Unix-like system with an SMT tool installed

References:

- ✓ Indic corpora : http://homepages.inf.ed.ac.uk/miles/babel.html
- ✓ Moses : http://www.statmt.org/moses/
- ✓ Joshua: http://joshua.incubator.apache.org
- ✓ Cdec: http://www.cdec-decoder.org/
- ✓ Sentence alignment: http://www.statmt.org/survey/Topic/SentenceAlignment
- ✓ Singh A.K., S Husain, 2005, Comparison, selection and use of sentence alignment algorithms for new language pairs, Proc. of the ACL Workshop on Building and Using Parallel Texts, 2005

Lecture 6: Evaluation / Quality (Bruno Pouliquen)

- Part 1: MT Evaluation
  - Human vs Automatic evaluation
  - Evaluating an SMT model with automatic metrics (BLEU, METEOR, other metrics)
- Part 2: Automatic Evaluation
  - Presentation of various metrics
  - Tuning a system (using Minimum Error Rate Training or MERT)
- Part 3: Quality estimation
  - Quality estimation: the process to learn a priori probabilities of machine translation
- Lab 6: Evaluation of MT quality
  - Evaluating a MT system. Use previously trained system, get a BLEU score. Compare various outputs from different systems.
  - Optimizing a system (MERT)

References:

  - Asiya (online MT evaluation tool) http://asiya.cs.upc.edu/demo/asiya_online.php
- M.R. Costa-jussà, M. Farrús, J.B. Mariño, J.A.R. Fonollosa. *Study and comparison of rule-based and statistical Catalan-Spanish machine translation systems*. Computing and Informatics, volume 31, issue 2, pages 245--270. February 2012. ISSN 1335-9150.
  - [lucia 2016] Lucia Specia "Translation quality assessment: Evaluation and Estimation", presentation at MTM 2016, Prague, "http://ufal.mff.cuni.cz/mtm16/files/01-mt-evaluation-and-quality-estimation-lucia-specia.pdf

Lecture 7: Neural Machine Translation (NMT) (Bruno Pouliquen)

- Part 1: Introduction to NMT
  - History
  - Recent developments (Google translate / Systran / WIPO etc.)
  - Overview of the method
  - Examples and comparison with PBSMT

- Part 2: How does it work
  - What are artificial neural networks?
  - Simple cases
  - Perceptron / Feed forward / Recurrent Neural networks (RNN)
- Part 3: Application to MT
  - RNN as a way to "learn" full translation process

References:

Tutorials:

Luong, Cho, Manning: https://sites.google.com/site/acl16nmt/
Rico Sennrich tutorial in MTM (Prag 2016): Also highly recommended
Mentions the BPE algorithm
http://ufal.mff.cuni.cz/mtm16/files/11-neural-machine-translation-rico-sennrich.pdf
A very nice bibliography at the end, please read!

Very good tutorial on NMT (highly recommended!):
http://nlp.stanford.edu/projects/nmt/Luong-Cho-Manning-NMT-ACL2016-v4.pdf
Very good tutorial about Neural Networks in general:
http://neuralnetworksanddeeplearning.com/

A video lecture about RNN (incl. image capture)
Fei-Fei Li & Andrej Karpathy & Justin Johnson
https://www.youtube.com/watch?v=iX5V1WpxxkY

Tensorflow playground :
https://github.com/tensorflow/playground

Lab 7: NMT (Bruno Pouliquen)
Use/modify a simple perceptron (language guessing)
Use an existing NMT model, compare output with PBSMT corresponding model
(Using Amun decoder, translate using a pre-trained NMT model)

References :

- https://github.com/rsennrich/nematus
- https://github.com/emjotde/amunmt

Items not covered
Transliteration:
"Improving Machine Translation via Triangulation and Transliteration" Durrani and Koehn (2014)

http://www.statmt.org/moses/?n=Advanced.OOVs