Taming Moses:

A practical introduction to Statistical Machine Translation

Carlos Ramisch

Porto Alegre - October 27, 2014 - CAMELEON Tutorial

Disclaimer

About this course

This course is intended as a general introduction to statistical machine translation for ***computer science*** and ***linguistics*** students and researchers. Some parts will seem ***extremely easy*** or ***extremely complicated*** according to your background. Please be patient and do not hesitate to ask questions at any time. ***There is no such thing as a silly question.***

Outline

Outline

What is SMT?

Statistical machine translation (SMT)

Learning a full translation system from text, using unsupervised statistical learning algorithms to find correspondences between units

A note on terminology

- Statistical vs Rule-based systems
- Statistical vs Example-based systems
- Probability vs Score
- SMT systems are learned empirically

Why is SMT so popular?

- Famous IBM papers [Brown et al., 1993]
- Freely available Moses toolkit
- Language pairs and parallel corpora
- Google translate

What is Moses?

- Moses is a ***translation toolkit***
- It includes an SMT decoder
- Free, open source
- Huge community of active developers
- Main maintainers: MT group of Univ. of Edinburgh

Further reading

- `http://www.statmt.org/moses/`
- Demo paper - [Koehn et al., 2007]
- Moses mailing list

Translation models (or paradigms)

- ***Phrase-based*** - units are sequences (most popular)
- ***Factored*** - units are multi-level sequences
- ***Hierarchical*** - units are "gappy" sequences
- ***Syntax-based*** - several flavours including trees

Preamble

- SMT is not a magic wand
- SMT is not cheap
- SMT requires tuning
- SMT is very useful
- SMT can quickly become a mess

Our goal

Learn how to use the Moses toolkit and related software (GIZA++, IRSTLM, mt-eval) to build a toy SMT system from scratch

Outline

XXX

- XXX

Outline

XXX

- XXX

Outline

N-gram language models in a nutshell

- Used in many applications as a proxy to syntax
  - Speech recognition
  - Text generation
  - Machine translation
  - ...
- Select a sentence s^* in the set S with maximal probability

$$s^* = \underset{s_j \in S}{\text{argmax}}\, p(s_j)$$

- The probability $p(s_j)$ is decomposed into n-grams

Mathematical formulation I

- Sentence s_j is a sequence of n words $w_1 \dots w_n$, noted w_1^n

$$p(s_j) = p(w_1^n) = p(w_1) \times p(w_2|w_1) \times p(w_3|w_1^2) \dots p(w_n|w_1^{n-1})$$

$$= p(w_1) \times \prod_{k=2}^{n} p(w_k|w_1^{k-1})$$

- Apply Markov's hypothesis: only short history determines the event

$$p(w_k|w_1^{k-1}) \approx p(w_k|w_{k-m+1}^{k-1})$$

- So that $p(s_j)$ is simplified as

$$p(w_1^n) = p(w_1) \times \prod_{k=2}^{n} p(w_k|w_{k-m+1}^{k-1}) = p(w_1) \times \prod_{k=2}^{n} \frac{p(w_{k-m+1}^{k})}{p(w_{k-m+1}^{k-1})}$$

Mathematical formulation II

- *N*-gram probabilities can be estimated through MLE (maximum likelihood estimator)

$$p(w_j^k) = \frac{c(w_j^k)}{N}$$

- Where N is the size of the corpus in tokens and $c(\cdot)$ is number of occurrences
- So the final formulation is

$$p(w_1^n) = \frac{c(w_1)}{N} \times \prod_{k=2}^{n} \frac{\frac{c(w_{k-m+1}^k)}{N}}{\frac{c(w_{k-m+1}^{k-1})}{N}} = \frac{c(w_1)}{N} \times \prod_{k=2}^{n} \frac{c(w_{k-m+1}^k)}{c(w_{k-m+1}^{k-1})}$$

Advanced language modelling

- Smoothing and discounting techniques
- Back-off strategies for OOV (out of vocabulary) *n*-grams when $c(w_i^j) = 0$
- Computationally costly
- Use of Bloom filters, recurrent neural networks, etc.

Some free tools

SRILM, IRSTLM, RandLM, KenLM, RNNLM

Exercices

1.

Outline

Example phrase table

Source s	**Target** t	$p(t\|s)$	$lex(t\|s)$	$p(s\|t)$	$lex(s\|t)$
a baby being born blind	uma criança cega	1	0.0106327	1	0.026239
a backward step .	de uma regressão .	1	0.0280532	0.5	0.002579
a backward step .	uma regressão .	1	0.0280532	0.5	0.027814
a backward step	de uma regressão	1	0.0287083	0.5	0.002676
a backward step	uma regressão	1	0.0287083	0.5	0.028855
a bad foundation for	uma má base para	1	0.0009332	1	0.004316
a bad foundation	uma má base	1	0.0036263	1	0.018618
a bad	uma má	1	0.1378	1	0.049648

XXX

- XXX

Exercices

1.

Outline

XXX

- XXX

Outline

XXX

- XXX

Merci beaucoup !
Muito obrigado!
Thank you!

Statistical machine translation

Carlos Ramisch (carlos.ramisch@lif.univ-mrs.fr)

References I

Brown, P. F., Pietra, S. D., Pietra, V. J. D., and Mercer, R. L. (1993).
The mathematics of statistical machine translation: Parameter estimation.
Comp. Ling.Special Issue on Using Large Corporа: II, 19(2):263–311.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007).
Moses: open source toolkit for statistical machine translation.
In *Proc. of the 45th ACL (ACL 2007)*, pages 177–180, Prague, Czech Republic. ACL.

APPENDIX