ACS Statistical Machine Translation

Lecture 6: Evaluating Machine Translation

Department of Engineering
University of Cambridge

Bill Byrne – bill.byrne@eng.cam.ac.uk

Lent 2013

Module L102: Statistical Machine Translation

- Lecture 1: Introduction to MT and SMT
- Lecture 2: Introduction to SMT Models
- Lecture 3: Word Alignment Models
- Lecture 4: Decoding with Phrase-Based Models
- Lecture 5: Introduction to Syntax-Based Models
- Lecture 6: MT Evaluation, Intro to Weighted Finite-State Automata
- Lecture 7: Phrase-based Translation with Weighted Finite-State Transducers
- Lecture 8: Hierarchical Phrase-based Translation
- Lecture 9: Hierarchical Translation Grammars, Syntax-based Translation
- Lecture 10: Minimum Error Training, Lattice Rescoring and System Combination
- Practical 1: Automatic Machine Translation Evaluation.
  Demonstrated Session: Monday 11th February 2013.
- Practical 2: Weighted Finite-State Transducers.
  Demonstrated Session: Monday 18th February 2013.
- Practical 3: Hierarchical Phrase-based Translation with alternative grammars
  Demonstrated Sessions: Monday 4th March 2013, Monday 11th March 2013.

Questions from Practical Handouts 1, 2 and 3 must be answered in a single report, which must be handed in by **18th March at 4.00pm**.

Machine Translation Quality

Machine translation quality is determined by the task it is expected to accomplish.

For a given task, humans must decide what is a 'good' or a 'bad' translation.

Proposed 'Human Metrics':

- A set of human judges assess each document/sentences/phrases
- Binary scores: correctness, task completion
- Scaled scores: fluency, adequacy
- Preference scores: which is better?
- Human-targeted or post-edit scores: minimal output correction $\leftarrow$ direct commercial impact

- $\times$ Slow and costly $\rightarrow$ cannot be used for extensive system development

Automatic Measurement of Translation Quality

Automatic performance metrics have been central to the development of large statistical language processing systems

- Word/Character Error Rate (WER): ASR , OCR, ...
- Precision/Recall: Information Retrieval, Speaker ID, ...
- Crossing Brackets: Parsing, ...

These are all relative to *human performance* over defined test sets

- human translations are obtained *once* over a fixed test set
- system performance can be measured *many times* relative to :
  - human performance, directly
  - performance of other systems, indirectly
- automatic metrics make incremental system improvement possible
- metrics can be incorporated into estimation and decoding algorithms

Developing automated evaluation metrics for Machine Translation is a research area

- *ACL 2008 Workshop on SMT, NIST 2008 MetricsMATR, EACL 2009 Workshop on SMT, ACL 2010 Workshop on SMT/MetricsMATR*

Evaluating MT. Automated metrics.

Automated evaluation metrics should:

- be inexpensive to compute
- require no human participation
- correlate with human perception of quality

- They compare system outputs against a set of golden reference translations
  - By comparing automatic translations to human references we obtain implicit measurements of fluency and accuracy

- A vast range of metrics have been proposed
  - At surface-form level: WER, PER, BLEU, NIST, TER, METEOR, GTM, ...
  - Incl. linguistic analysis: POS-BLEU, TERP, ULC, ...

- In general, they are pessimistic but we expect them to be useful for ranking purposes
  - There may be more correct ways of translating a text than the ones shown in the references

- The number of available references is important

Automatic Metrics for MT (1)

WER is standard Word Error Rate:

$$WER(T,R) = \frac{Ins + Del + Sub}{N} \times 100\%$$

- N is the (average) number of words in the reference
- Standard in speech recognition evaluation, where a single reference applies
- Allows insertions, deletions and substitutions with equal cost
- Adapted to multiple references for MT

TER [1] is the Translation Edit Rate, defined as:

$$TER(T,R) = \frac{Ins + Del + Sub + Shft}{N} \times 100\%$$

- Extends WER with shifts to account for reordering
- Shift is the movement of a continuous block from the hypothesis

[1] Snover, M. et al. 2006. A study of translation edit rate with targeted human annotation, Proc. AMTA

Automatic Metrics for MT (2)

BLEU[2] is an MT metric based on n-gram precision

- An example of computing Bleu against a single reference translation:

Reference : mr. speaker , in absolutely no way .
Hypothesis : in absolutely no way , mr. chairman .

BLEU Computation

n-gram matches				BLEU
1-word	2-word	3-word	4-word	$\left(\frac{7}{8} \times \frac{3}{7} \times \frac{2}{6} \times \frac{1}{5}\right)^{\frac{1}{4}} = 0.3976$
7/8	3/7	2/6	1/5	

- Can be easily generalized to multiple references
- Also includes a length penalty (penalises short translation hypotheses)
- Correlates well with human judgments of translation

N-grams precisions are computed accross multiple sentences in a set

Not so adequate on a sentence level (as in example above)

[2] Papineni, K., Roukos, S., Ward, T., & Zhu, W. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. Pages 311–318 of: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).

Automatic metrics for MT (3)

NIST [3] is a variant of BLEU

- Assigns different value to each matching n-gram according to information gain statistics
- Less sensitive to brevity penalty ('gaussian' length distribution near average ref length)
- Score ranges from 0 (worst translation) to an unlimited positive value

METEOR [4] is harmonic mean of unigram precision and recall:

$$METEOR = \frac{10PR}{R + 9P} \times (1 - p)$$

- includes penalty $p \propto$ number of alignment chunks between hyp and refs
- accepts synonyms and considers stemming, but WordNet has to be available

Others:

- ROUGE and ORANGE use techniques from summarization evaluation
- Geometric Translation Mean (GTM), Weighted N-gram Model (WNM), Classification Error Rate (CER), ...

  $\Rightarrow$ No metric has overcome BLEU in terms of widespread use yet

[3] Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, Proc. ARPA Workshop on HLT.

[4] Banerjee,S. and Lavie,A. 2005. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments", Proc. of the ACL Workshop onEvaluation Measures for MT.

Calculation of BLEU

The goal is to calculate the BLEU score of a set of automatic translations $\{E^i\}_{i=1}^S$ against a set of reference translations, e.g. $\{E^i_{(1)}, E^i_{(2)}, E^i_{(3)}, E^i_{(4)}\}_{i=1}^S$.

- Set N to be the order of the highest n-gram to be considered, e.g. N=4
- For each sentence i, and for $n = 1, \ldots, N$, gather the following n-gram counts:
  - c_n^i : the number of hypothesized n-grams
  - $\bar{c}_n^i$: the number of correct n-grams, where the contribution of each distinct n-gram is *clipped* to the maximum number of occurrences in any one reference

  Example:

  Hypothesis: the the the the the the the
  Reference 1: the cat is on the mat
  Reference 2: there is a cat on the mat

  In this example, $c_1 = 7$, but $\bar{c}_1 = 2$
- Compute the precision for each n-gram , $n = 1, \ldots, N$: $p_n = \left(\sum_i \bar{c}_n^i\right) / \left(\sum_i c_n^i\right)$
- Calculate the Brevity Penalty
  - Compute the hypothesis length : $l = \sum_i |E^i|$
  - Compute the closest reference length : $r = \sum_i \min_r \{|E^i_{(r)} - E^i|\}$

$$BP = min\left\{1, \frac{l}{r}\right\}$$

- The BLEU score is

$$\text{BLEU} = BP * \exp\{\sum_{n=1}^{N} \frac{1}{N} \log p_n\}$$

BLEU Assigns High Scores to Good Translations

Example translations at various levels of translation performance as measured by the sentence-level BLEU score.

Translations	BLEU (%)
	60 – 70%
Afghan Earthquake Victims begin to rebuild their homes .	66.1
Prior to this , the ANC has issued a statement calling for the international community to respect the choice of the people and help them survive .	66.0
Statistics show that since 1992 , a total of 204 UN personnel have been killed , but only 15 criminals have been arrested .	64.4
Chavez emphasized that Venezuela needs peace , stability and reason for all parties should make joint efforts to end the conflict .	62.2
London Financial Times Index Friday at closing newspaper 5,292.70 points , up 31.30 points .	61.0

readable and fairly plausible

BLEU Assigns Middling Scores to Middling Translations

	$20-30\%$
Japan to temporarily freeze asked Russia to provide humanitarian assistance ,	30.0
Opposition Senator held that the president should focus more on domestic affairs and not eager to go abroad .	26.2
Taiwan DPP Legislator Chen Kim de fisheries groups to visit to Beijing .	23.9
Recently , the international community for the recent conflict , the fiercest Jenin camp conflict investigation of spreading .	20.8
opinion maintained : Gusmao victory is a strong possibility because he is considered the East Timor independence hero .	20.0

- probably misleading with respect to details
- contains readable sections

BLEU Assigns Low Scores to Poor Translations

	$0-10\%$
Japan Telecom company in 2000 to spend 5.5 billion dollars buy back .	0.0
However , the voting result shows that Zhu because there is no reason to be losing power by NPC deputies desolate .	0.0
77 private manufacturing enterprises also reported a foreign trade management right .	0.0
Identification Department found that college students of the certificate , many of them were fake .	0.0
The European Union would be implemented in steel imports temporary protective measures to discuss with the Chinese side ,	0.0
Georgia from a section of the great mountains Canyon withdrawal ,	0.0

barely readable and probably misleading

An Example of Good (but not perfect) Translation

Original Chinese Gloss

(By) (2005 year) (internet) (,) (whole country) (users) (will) (reach) (0.2 billion)

Automatic Translation

By 2005 , the number of internet users will reach 200 million

Human Reference Translations

By 2005 , the number of internet users in the whole country will reach 200 million
By 2005 , the number of internet users in China is estimated to be 200 million
By 2005 , internet customers across the country is to reach 200 million
In 2005 , the internet users in China will total 0.2 billion

- The translation agrees fairly closely with the reference translations
- The translation is close to fluent
- Note the variability in the four reference translations.

An Example of Relatively Poor Translation

Original Word Segmented Chinese Pinyin (with tones) and gloss

sui1ran2 bei3feng1 hu1xiao4 , dan4 tian1kong1 yi1ran2 shi2fen1 qing1che4
(Although) (northern wind) (howl) (,) (but) (sky) (very) (clear)

Automatic Translation

Although wind howl . but the skies remain very tender

Human References

Although a north wind was howling , the sky remained clear and blue .
However , the sky remained clear under the strong north wind .
Despite the strong northerly winds , the sky remains very clear .
The sky was still crystal clear , though the north wind was howling .

- some resemblance to the reference translations
- not fluent
- questionable accuracy

Reference translation variability is a large problem...

... which gets worse as the domains get more interesting

BLEU can be used for SMT development

BLEU is not an absolute measure of translation performance

- unlike Word Error Rate used in speech recognition
- most useful in indicating the relative quality between two different systems

Assessment using BLEU can be trusted when :

- many different translation systems are developed under BLEU, and results throughout development are published on standard test sets
- performance is frequently validated against human judgments

Current test sets used in the NIST MT evaluations

- $\sim 1K$ sentences / $\sim 25K$ words,
- four independently produced reference translations

There is extensive effort in improving/extending/replacing BLEU

Practical 1/3

- Automatic Evaluation of Machine Translation using BLEU
- Handout available at:

  `http://www.cl.cam.ac.uk/teaching/1213/L102/materials.html`
- Demonstrated Session: 11th February
- Answers to practical questions should be included in a single practical report to be handed at the end of term

Weighted Finite-State Transducers (WFSTs)

- General framework of structures/algorithms
  useful to encode/process conditional probability distributions

- Used in many Speech and Natural Language Processing tasks

- ▷ Well-suited for carrying out search procedures involving Markov processes and HMMs
- ▷ Efficient, standard algorithms can be applied directly

Introduction

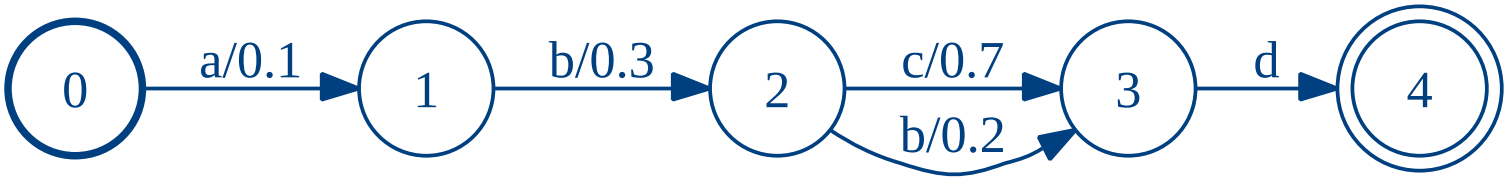
Weighted Finite State Automata (WFSA):

- General framework of structures/algorithms that are useful to encode and process conditional probability distributions
- Used in many Speech and Natural Language Processing tasks
- Well-suited to carry out search procedures involving Markov processes and HMMs
- If a problem can be cast in a WFSA framework, efficient standard algorithms can be applied directly

Intro: Finite State Automata - Weighted Acceptors

Weighted acceptors can assign costs to strings
- weights (or costs) are accumulated over **paths** through the automata
- a path is a sequence of edges (or arcs) associated with a string

A weighted automaton which accepts only two strings:



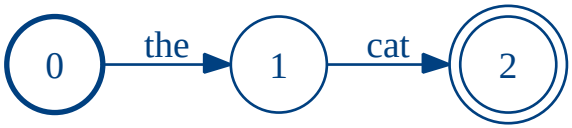
$$w(\text{'a b c d'}) = 0.1 + 0.3 + 0.7 + 0.0 = 1.1$$
$$w(\text{'a b b d'}) = 0.1 + 0.3 + 0.2 + 0.0 = 0.6$$

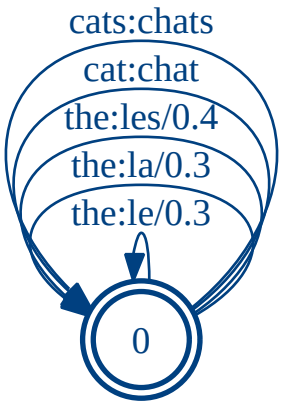
To define the acceptor, we specify a set of states Q and a set of arcs : $q \stackrel{x/k}{\rightarrow} q'$
- q is the start state, q' is the end state, x is the input symbol, k is the arc weight

Semirring: weights can either be probabilities, negative log-probabilities, etc.

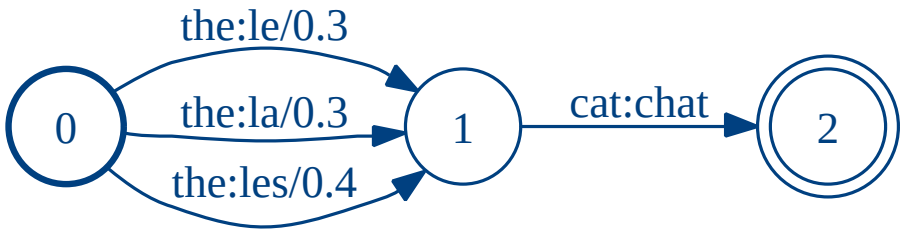
Intro: Finite State Automata - Weighted Transducers

An acceptor and a weighted transducer:





Their **composition** (more on this later):



the cat : le chat / 0.3
the cat : la chat / 0.3
the cat : les chat / 0.4

To describe the transducer, we specify a set of states and a set of arcs : $q \overset{x:y/k}{\rightarrow} q'$
- x is the input symbol, y is the output symbol, k is the arc weight

Assigning weights to paths is useful to describe ambiguity and uncertainty

Weighted Finite State Acceptors - Definition

A weighted acceptor over a finite input alphabet Σ is a finite directed graph with a set of nodes Q (states) and a set of arcs E (edges).
- Each arc (or edge) e has an initial (or start) state $s(e)$ and a final state $f(e)$.
- Each arc e is labeled with an input symbol $i(e)$ and a weight $w(e)$.
- The weights take values in $\mathbb{K}$

A complete **path** through an acceptor can be written as $p = e_1 \cdots e_{n_p}$, where :
- the path p consists of n_p edges
- the path starts at state $i_p = s(e_1)$ where i_p is an initial state
- the path ends at state $f_p = f(e_{n_p})$ where f_p is a final state

The arc weights and initial and final weights combine to form the **path weight**

$$w(p) = \lambda(i_p) \otimes w(e_1) \otimes \cdots \otimes w(e_{n_p}) \otimes \rho(f_p)$$

- Initial weights and final weights : $\lambda(i_p)$ and $\rho(f_p)$
- $\otimes$ is the **product** of two weights (to be defined shortly)

Notation: $\otimes_{j=1}^{n_p} w(e_j) = w(e_1) \otimes \cdots \otimes w(e_{n_p})$ so that $w(p) = \lambda(i_p) \otimes (\otimes_{j=1}^{n_p} w(e_j)) \otimes \rho(f_p)$

Weights Assigned to Strings by Acceptors

Since each arc in the WFSA has an input symbol, it is straightforward to associate paths through the acceptor with input sequences.

- A path $p = e_1 \cdots e_{n_p}$ produces the string $x = i(e_1) \cdots i(e_{n_p})$

If every string was generated by a unique path through an acceptor, assigning weights to strings would be easy: the string weight would be its path weight. However, since strings can be generated by multiple paths, the acceptor combines the weights of all paths which might have generated a string, as follows:

- Let x be a string constructed from symbols in the input alphabet Σ : $x \in \Sigma^*$
- Let $P(x)$ be the set of complete paths which generate x, i.e. $x = i(e_1) \cdots i(e_{n_p})$
- Let $\oplus$ be the *sum* of two weight values
- Define $[\![A]\!](x)$ as the cost assigned to the string x by the transducer

$$[\![A]\!](x) = \bigoplus_{p \in P(x)} \underbrace{\lambda(i_p) \otimes (\otimes_{j=1}^{n_p} w(e_j)) \otimes \rho(f_p)}_{w(p)}$$

$[\![A]\!](x)$ is the 'Sum' of the weights of the complete paths which can generate x

Weights and Operations on Weights

The *product* operation $\otimes$ is used to compute the weight of a single path from the weights of its edges

The *sum* operation $\oplus$ is used to compute the weight of a sequence by summing over all the distinct paths which could have generated that sequence

Semirings : sum $\oplus$ and product $\otimes$ with identity elements $\bar{0}$ and $\bar{1}$
- For a weight $k \in \mathbb{K}$: $\bar{0} \oplus k = k$; $\bar{1} \otimes k = k$; $\bar{0} \otimes k = \bar{0}$
- $\oplus$ and $\otimes$ distribute and commute in the familiar way

Three useful semirings:

Semiring	$\mathbb{K}$	$\oplus$	$\otimes$	$\bar{0}$	$\bar{1}$
Probability	$\mathbb{R}_+$	$+$	$\times$	0	1
Log	$\mathbb{R} \cup \{-\infty, \infty\}$	$\oplus_{\log}$	$+$	∞	0
Tropical	$\mathbb{R} \cup \{-\infty, \infty\}$	$\min$	$+$	∞	0

$$\oplus_{\log} : k_1 \oplus_{\log} k_2 = -\log(e^{-k_1} + e^{-k_2})$$

Unless otherwise stated, the tropical semiring is used by default

WFSAs and N-gram Language Models

WFSAs can be used to implement N-Gram language models. Back-off N-Gram language models can be encoded using failure transitions (taken if no other arc can be taken). Recall the back-off bigram language model:

$$\hat{P}(w_j|w_i) = \begin{cases} p(w_i, w_j) & f(w_i, w_j) > C \\ \alpha(w_i)\hat{P}(w_j) & \text{otherwise} \end{cases}$$

where $p(w_i, w_j) = d(f(w_i, w_j))\frac{f(w_i,w_j)}{f(w_i)}$.

Language model vocabulary : $\Sigma = \{a, b\}$

Cutoff statistics :

$f(a, b) > C$

but $f(b, a) < C$

Bigram probabilities:

$P(b|a) = p(a, b) \leftarrow$ no back-off

$P(a|b) = \alpha(b)\hat{P}(a) \leftarrow$ back-off