Introduction to
the Evaluation of Machine Translation

slides from François Yvon
with contributions from the SMT@LIMSI crew

LIMSI-CNRS & Université Paris Sud

ELDA, Paris | November 29, 2012

Disclaimer

Evaluation of MT : a field in itself

« More has been written about MT evaluation over the past 50 years than about MT itself » [1], [2]
« it is impossible to write a comprehensive overview of the MT evaluation literature » [1]

[1] Eduard Hovy, Margaret King et Andrei Popescu-Belis, Principles of Context-Based Machine Translation Evaluation. Machine Translation, 16, pp. 1–33, 2002

[2] Statistical Machine Translation, Adam Lopez, In ACM Computing Surveys 40(3) pp. 1–49, August 2008.

Human judgments

Adequacy and *Fluency*

- *adequacy* : is meaning preserved ?
- *fluency* : is the translation fluent ?
- scores on some scale (usually, between 1 and 5)

A controversial approach

« *Low inter-judge correlation underscores how little the community understands about the MT evaluation problem. If the MT research community is serious about designing reliable automatic MT evaluation measures, then we must obtain human judgment data through more reliable means.* » [3]

[3] Joseph P. Turian, Luke Shen, and I. Dan Melamed. Evaluation of machine translation and its evaluation. In Proc. MT Summit IX 2003.

Human judgments

Adequacy and *Fluency*

- *adequacy* : is meaning preserved ?
- *fluency* : is the translation fluent ?
- scores on some scale (usually, between 1 and 5)

A controversial approach

« *Low inter-judge correlation underscores how little the community understands about the MT evaluation problem. If the MT research community is serious about designing reliable automatic MT evaluation measures, then we must obtain human judgment data through more reliable means.* » [3]

[3] Joseph P. Turian, Luke Shen, and I. Dan Melamed. Evaluation of machine translation and its evaluation. In **P**roc. MT Summit IX 2003.

Input text: C'était très agréable, quand je restais tard en ville , de me lancer dans la nuit, surtout si elle était sombre et orageuse , et de faire voile depuis quelque lumineux salon du village ou de quelque salle de conférence, un sac de farine de seigle ou de maïs sur l'épaule, vers mon port bien abrité dans les bois, ayant tout bien fermé et m'étant retiré derrière les écoutilles avec un joyeux équipage de pensées, ne laissant que mon homme de surface à la barre, ou même attachant le gouvernail quand tout allait bien.

Candidate translation: It was very agreeable, when I remained in town, late in the night, especially if she was dark and stormy, and to make sail for some luminous drawing-room hall of some of the village, or to conference, a bag of flour of rye and Indian corn, on the shoulder to my well sheltered harbour in the woods, having all well shut behind me and being removed with the thoughts of a happy crew, leaving my surface of that man at the wheel, tying the rudder, or even when all was well.

Input text: C’était très agréable, quand je restais tard en ville , de me lancer dans la nuit, surtout si elle était sombre et orageuse , et de faire voile depuis quelque lumineux salon du village ou de quelque salle de conférence, un sac de farine de seigle ou de maïs sur l’épaule, vers mon port bien abrité dans les bois, ayant tout bien fermé et m’étant retiré derrière les écoutilles avec un joyeux équipage de pensées, ne laissant que mon homme de surface à la barre, ou même attachant le gouvernail quand tout allait bien.

Candidate translation: It was very agreeable, when I remained in town, late in the night, especially if she was dark and stormy, and to make sail for some luminous drawing-room hall of some of the village, or to conference, a bag of flour of rye and Indian corn, on the shoulder to my well sheltered harbour in the woods, having all well shut behind me and being removed with the thoughts of a happy crew, leaving my surface of that man at the wheel, tying the rudder, or even when all was well.

Automatic evaluation metrics

Requirements

- compare systems on a given translation task
- compare variants of a given system
- assist in performing error analysis
- should not be too expensive

Issues

- many acceptable (and unacceptable) translations
- evaluation (*adequacy*) is as difficult as understanding
- evaluation (*fluency*) requires high robustness

Automatic evaluation metrics

Requirements

- compare systems on a given translation task
- compare variants of a given system
- assist in performing error analysis
- should not be too expensive

Issues

- many acceptable (and unacceptable) translations
- evaluation (*adequacy*) is as difficult as understanding
- evaluation (*fluency*) requires high robustness

Statistical Machine Translation: the noisy channel model

Noisy channel

Given $\mathbf{f}$ a source language sentence, translating is equivalent to solving :

$$\mathbf{e}^* = \operatorname*{argmax}_{\mathbf{e}} P(\mathbf{e}|\mathbf{f}) = \operatorname*{argmax}_{\mathbf{e}} P(\mathbf{f}|\mathbf{e})P(\mathbf{e})$$

where the maximum is found over the set of all sentences $\mathbf{e}$ in the target language

Two important problems for Statistical MT

- define and estimate the probabilistic models
- solve the optimization problem

Statistical Machine Translation: the noisy channel model

Noisy channel

Given $\mathbf{f}$ a source language sentence, translating is equivalent to solving :

$$\mathbf{e}^* = \operatorname*{argmax}_{\mathbf{e}} P(\mathbf{e}|\mathbf{f}) = \operatorname*{argmax}_{\mathbf{e}} P(\mathbf{f}|\mathbf{e})P(\mathbf{e})$$

where the maximum is found over the set of all sentences $\mathbf{e}$ in the target language

Two important problems for Statistical MT

- define and estimate the probabilistic models
- solve the optimization problem

Probabilistic models for translation

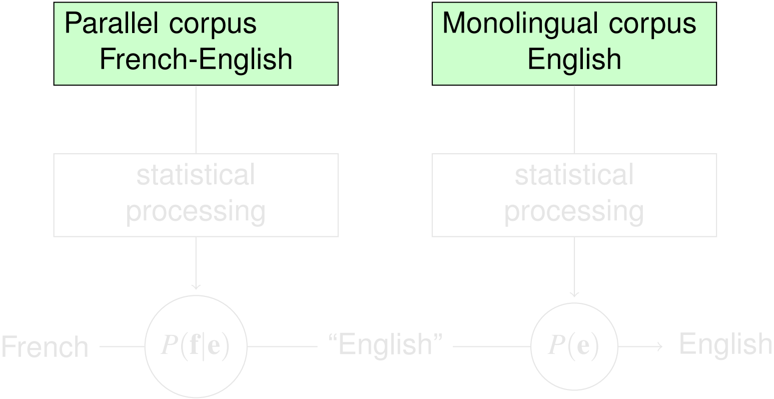
$P(\mathbf{f}|\mathbf{e})$ = translation model

- measure the quality of the pairing between $\mathbf{e}$ and $\mathbf{f}$
- estimated from (large) parallel corpora

$P(\mathbf{e})$ = language model

- measure some notion of quality of a translation candidate
- estimated from (very large) monolingual corpora

Resources, models, algorithms



Parallel corpus
French-English

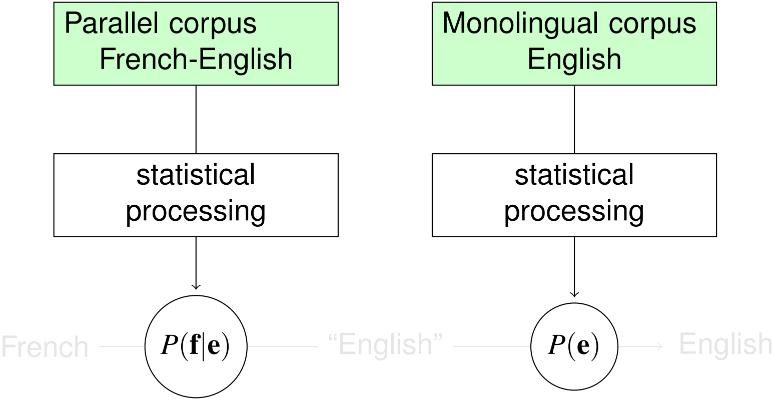
Monolingual corpus
English

statistical
processing

statistical
processing

French $P(\mathbf{f}|\mathbf{e})$ “English” $P(\mathbf{e})$ English

Resources, models, algorithms



Parallel corpus French-English

statistical processing

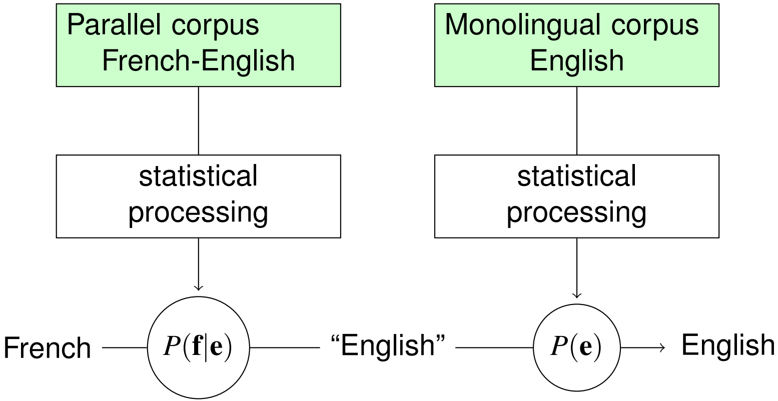
$P(\mathbf{f}|\mathbf{e})$

Monolingual corpus English

statistical processing

$P(\mathbf{e})$

Resources, models, algorithms



Parallel corpus French-English

Monolingual corpus English

statistical processing

statistical processing

French

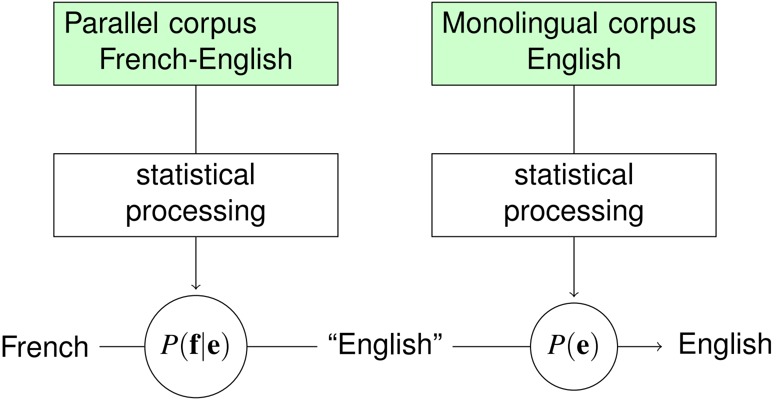
$P(\mathbf{f}|\mathbf{e})$

"English"

$P(\mathbf{e})$

English

Resources, models, algorithms



Parallel corpus French-English

Monolingual corpus English

statistical processing

statistical processing

French

$P(\mathbf{f}|\mathbf{e})$

“English”

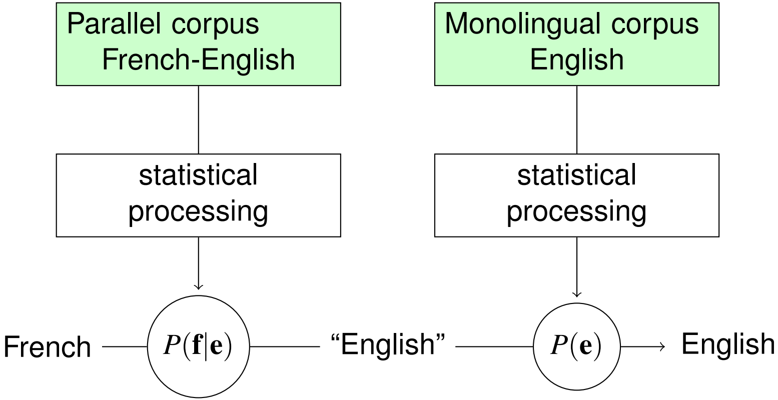
$P(\mathbf{e})$

English

Decoding:

$$\text{argmax } P(\mathbf{f}|\mathbf{e})P(\mathbf{e})$$

Resources, models, algorithms



Decoding:

$$\text{argmax } \log(P(\mathbf{f}|\mathbf{e})) + \lambda \log(P(\mathbf{e}))$$

Phrase-based Statistical MT

$$
\begin{aligned}
P(\mathbf{f}|\mathbf{e}) &= \sum_{f_1\ldots f_K=\mathbf{f}, e_1\ldots e_K=\mathbf{e}} \prod_{k=1}^{K} P(\mathbf{f}_k|\mathbf{e}_k) \\
&\approx \max_{f_1\ldots f_K=\mathbf{f}, e_1\ldots e_K=\mathbf{e}} \prod_{k=1}^{K} P(\mathbf{f}_k|\mathbf{e}_k)
\end{aligned}
$$

(official forecasts) (predicted) (just 3 percent) (, Bloomberg said)
(les prévisions officielles) (prévoyait) (seulement 3 %) (, a dit Bloomberg)

- capture local context
- capture some local reorderings
- capture idioms and terms

Phrase-based Statistical MT

$$
\begin{aligned}
P(\mathbf{f}|\mathbf{e}) &= \sum_{f_1 \ldots f_K = \mathbf{f}, e_1 \ldots e_K = \mathbf{e}} \prod_{k=1}^{K} P(\mathbf{f}_k|\mathbf{e}_k) \\
&\approx \max_{f_1 \ldots f_K = \mathbf{f}, e_1 \ldots e_K = \mathbf{e}} \prod_{k=1}^{K} P(\mathbf{f}_k|\mathbf{e}_k)
\end{aligned}
$$

(official forecasts) (predicted) (just 3 percent) (, Bloomberg said)
(les prévisions officielles) (prévoyait) (seulement 3 %) (, a dit Bloomberg)

- capture local context
- capture some local reorderings
- capture idioms and terms

Phrase-based Statistical MT

$$
\begin{aligned}
P(\mathbf{f}|\mathbf{e}) &= \sum_{f_1 \dots f_K = \mathbf{f}, e_1 \dots e_K = \mathbf{e}} \prod_{k=1}^{K} P(\mathbf{f}_k|\mathbf{e}_k) \\
&\approx \max_{f_1 \dots f_K = \mathbf{f}, e_1 \dots e_K = \mathbf{e}} \prod_{k=1}^{K} P(\mathbf{f}_k|\mathbf{e}_k)
\end{aligned}
$$

(official forecasts) (predicted) (just 3 percent) (, Bloomberg said)
(les prévisions officielles) (prévoyait) (seulement 3 %) (, a dit Bloomberg)

- capture local context
- capture some local reorderings
- capture idioms and terms

Decoding in the standard model in practice

From the noisy channel...

$$\mathbf{e}^* = \operatorname{argmax} \log(P(\mathbf{f}|\mathbf{e})) + \lambda \log(P(\mathbf{e}))$$

... to the *log-linear* model

$$\mathbf{e}^* = \operatorname{argmax} \sum_k \lambda_k F_k(\mathbf{f}, \mathbf{e}, \mathbf{a})$$

- F_ks evaluate various aspects of the association between $\mathbf{f}$ and $\mathbf{e}$

Decoding in the standard model in practice

From the noisy channel...

$$\mathbf{e}^* = \operatorname{argmax} \log(P(\mathbf{f}|\mathbf{e})) + \lambda \log(P(\mathbf{e}))$$

... to the *log-linear* model

$$\mathbf{e}^* = \operatorname{argmax} \sum_k \lambda_k F_k(\mathbf{f}, \mathbf{e}, \mathbf{a})$$

- F_ks evaluate various aspects of the association between $\mathbf{f}$ and $\mathbf{e}$

Various aspects of a translation

$s(\mathbf{f}, \mathbf{e}) = \sum_k \lambda_k F_k(\mathbf{f}, \mathbf{e}, \mathbf{a})$

- $F_k(\mathbf{f}, \mathbf{e}, \mathbf{a}) = \log P(\mathbf{e})$: language model
- $F_k(\mathbf{f}, \mathbf{e}, \mathbf{a}) = \sum_i \log P(\mathbf{e}_i | \mathbf{f}_i)$: translation model
- $F_k(\mathbf{f}, \mathbf{e}, \mathbf{a}) = \sum_i \log P(\mathbf{f}_i | \mathbf{e}_i)$: inverse translation model
- $F_k(\mathbf{f}, \mathbf{e}, \mathbf{a}) = \sum_i \log P_w(\mathbf{f}_i | \mathbf{e}_i)$ word alignment model
- reordering models
- possibly many other models (syntactic, etc.)

Issue : automatic learning of λ_k

Various aspects of a translation

$s(\mathbf{f}, \mathbf{e}) = \sum_k \lambda_k F_k(\mathbf{f}, \mathbf{e}, \mathbf{a})$

- $F_k(\mathbf{f}, \mathbf{e}, \mathbf{a}) = \log P(\mathbf{e})$: language model
- $F_k(\mathbf{f}, \mathbf{e}, \mathbf{a}) = \sum_i \log P(\mathbf{e}_i | \mathbf{f}_i)$: translation model
- $F_k(\mathbf{f}, \mathbf{e}, \mathbf{a}) = \sum_i \log P(\mathbf{f}_i | \mathbf{e}_i)$: inverse translation model
- $F_k(\mathbf{f}, \mathbf{e}, \mathbf{a}) = \sum_i \log P_w(\mathbf{f}_i | \mathbf{e}_i)$ word alignment model
- reordering models
- possibly many other models (syntactic, etc.)

Issue : automatic learning of λ_k

Decoding: a simplified example (courtesy of Ph. Langlais)

input: *Un bon avocat de Perpignan défendra cet escroc*

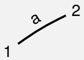
1

Translation table		
un	↔	a
un bon	↔	a good
	↔	a brilliant
	↔	a tasty
	↔	some good
bon	↔	good
	↔	brilliant
avocat	↔	lawyer
	↔	avocado
de	↔	of
	↔	from
Perpignan	↔	Perpignan
défendra	↔	will defend
cet escroc	↔	this crook

Language model	
a good lawyer	:-)
a brilliant lawyer	:-|
a tasty lawyer	:-(
a good avocado	:-|
a brilliant avocado	:-(
a tasty avocado	:-)
lawyer from Perpignan	:-(
avocado from Perpignan	:-(

Decoding: a simplified example (courtesy of Ph. Langlais)

input: *Un bon avocat de Perpignan défendra cet escroc*

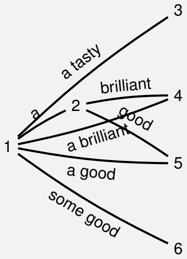


Translation table		
un	↔	a
un bon	↔	a good
	↔	a brilliant
	↔	a tasty
	↔	some good
bon	↔	good
	↔	brilliant
avocat	↔	lawyer
	↔	avocado
de	↔	of
	↔	from
Perpignan	↔	Perpignan
défendra	↔	will defend
cet escroc	↔	this crook

Language model	
a good lawyer	:-)
a brilliant lawyer	:-|
a tasty lawyer	:-(
a good avocado	:-|
a brilliant avocado	:-(
a tasty avocado	:-)
lawyer from Perpignan	:-(
avocado from Perpignan	:-(

Decoding: a simplified example (courtesy of Ph. Langlais)

input: *Un bon avocat de Perpignan défendra cet escroc*



Translation table

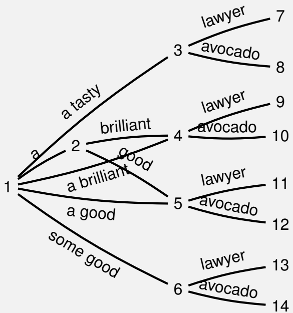
un	↔	a
un bon	↔	a good
	↔	a brilliant
	↔	a tasty
	↔	some good
bon	↔	good
	↔	brilliant
avocat	↔	lawyer
	↔	avocado
de	↔	of
	↔	from
Perpignan	↔	Perpignan
défendra	↔	will defend
cet escroc	↔	this crook

Language model

a good lawyer	:-)
a brilliant lawyer	:-|
a tasty lawyer	:-(
a good avocado	:-|
a brilliant avocado	:-(
a tasty avocado	:-)
lawyer from Perpignan	:-(
avocado from Perpignan	:-(

Decoding: a simplified example (courtesy of Ph. Langlais)

input: *Un bon avocat de Perpignan défendra cet escroc*

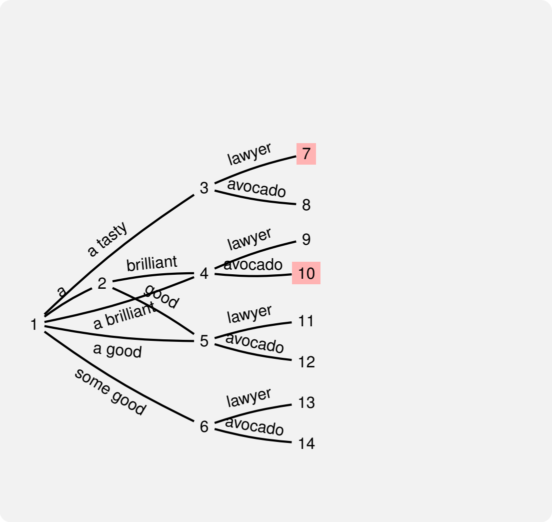


Translation table		
un	↔	a
un bon	↔	a good
	↔	a brilliant
	↔	a tasty
	↔	some good
bon	↔	good
	↔	brilliant
avocat	↔	lawyer
	↔	avocado
de	↔	of
	↔	from
Perpignan	↔	Perpignan
défendra	↔	will defend
cet escroc	↔	this crook

Language model	
a good lawyer	:-)
a brilliant lawyer	:-|
a tasty lawyer	:-(
a good avocado	:-|
a brilliant avocado	:-(
a tasty avocado	:-)
lawyer from Perpignan	:-(
avocado from Perpignan	:-(

Decoding: a simplified example (courtesy of Ph. Langlais)

input: *Un bon avocat de Perpignan défendra cet escroc*

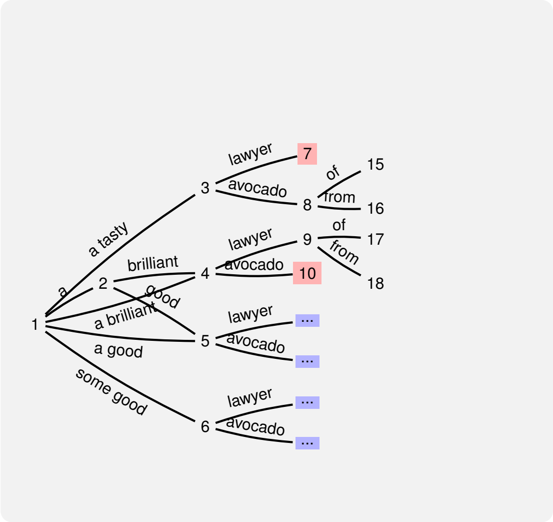


Translation table		
un	↔	a
un bon	↔	a good
	↔	a brilliant
	↔	a tasty
	↔	some good
bon	↔	good
	↔	brilliant
avocat	↔	lawyer
	↔	avocado
de	↔	of
	↔	from
Perpignan	↔	Perpignan
défendra	↔	will defend
cet escroc	↔	this crook

Language model	
a good lawyer	:-)
a brilliant lawyer	:-|
a tasty lawyer	:-(
a good avocado	:-|
a brilliant avocado	:-(
a tasty avocado	:-)
lawyer from Perpignan	:-(
avocado from Perpignan	:-(

Decoding: a simplified example (courtesy of Ph. Langlais)

input: *Un bon avocat de Perpignan défendra cet escroc*

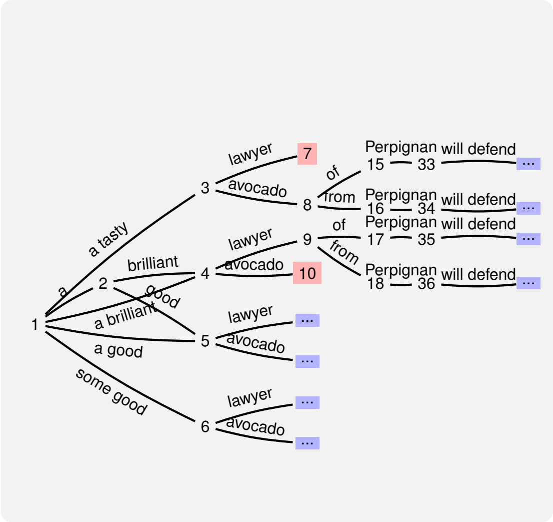


Translation table		
un	↔	a
un bon	↔	a good
	↔	a brilliant
	↔	a tasty
	↔	some good
bon	↔	good
	↔	brilliant
avocat	↔	lawyer
	↔	avocado
de	↔	of
	↔	from
Perpignan	↔	Perpignan
défendra	↔	will defend
cet escroc	↔	this crook

Language model	
a good lawyer	:-)
a brilliant lawyer	:-|
a tasty lawyer	:-(
a good avocado	:-|
a brilliant avocado	:-(
a tasty avocado	:-)
lawyer from Perpignan	:-(
avocado from Perpignan	:-(

Decoding: a simplified example (courtesy of Ph. Langlais)

input: *Un bon avocat de Perpignan défendra cet escroc*



Translation table

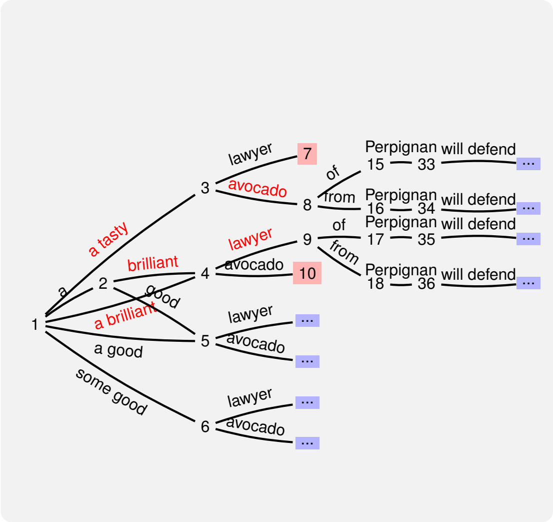
un	↔	a
un bon	↔	a good
	↔	a brilliant
	↔	a tasty
	↔	some good
bon	↔	good
	↔	brilliant
avocat	↔	lawyer
	↔	avocado
de	↔	of
	↔	from
Perpignan	↔	Perpignan
défendra	↔	will defend
cet escroc	↔	this crook

Language model

a good lawyer	:-)
a brilliant lawyer	:-|
a tasty lawyer	:-(
a good avocado	:-|
a brilliant avocado	:-(
a tasty avocado	:-)
lawyer from Perpignan	:-(
avocado from Perpignan	:-(

Decoding: a simplified example (courtesy of Ph. Langlais)

input: *Un bon avocat de Perpignan défendra cet escroc*



Translation table		
un	↔	a
un bon	↔	a good
	↔	a brilliant
	↔	a tasty
	↔	some good
bon	↔	good
	↔	brilliant
avocat	↔	lawyer
	↔	avocado
de	↔	of
	↔	from
Perpignan	↔	Perpignan
défendra	↔	will defend
cet escroc	↔	this crook

Language model	
a good lawyer	:-)
a brilliant lawyer	:-|
a tasty lawyer	:-(
a good avocado	:-|
a brilliant avocado	:-(
a tasty avocado	:-)
lawyer from Perpignan	:-(
avocado from Perpignan	:-(

Development cycle

1. estimate models independently on training data
2. adjust λ_k on tuning data by optimizing on quality

Minimum error rate training

$$s(H,R) = \sum_k \lambda_k F_k(H,R,\mathbf{a})$$

1. compute n-best hypotheses with fixed λ: $\{H_1 \ldots H_n\}$ s.t. $s(H_1,\lambda) > s(H_2,\lambda) > ... > s(H_n,\lambda)$
2. evaluation these n hypotheses $q(H_1,R) \ldots q(H_n,R)$
3. adjust λ s.t. $q(H_i,R) > q(H_j,R) \Rightarrow s(H_i,\lambda) > s(H_j,\lambda)$
4. convergence ? or back to 1

⇒ difficult optimization problem

⇒ requires suitable automatic metrics

Development cycle

1. estimate models independently on training data
2. adjust λ_k on tuning data by optimizing on quality

Minimum error rate training

$$s(H,R) = \sum_k \lambda_k F_k(H,R,\mathbf{a})$$

1. compute n-best hypotheses with fixed λ:
   $\{H_1 \ldots H_n\}$ s.t. $s(H_1,\lambda) > s(H_2,\lambda) > ... > s(H_n,\lambda)$
2. evaluation these n hypotheses $q(H_1,R) \ldots q(H_n,R)$
3. adjust λ s.t. $q(H_i,R) > q(H_j,R) \Rightarrow s(H_i,\lambda) > s(H_j,\lambda)$
4. convergence ? or back to 1

⇒ difficult optimization problem

⇒ requires suitable automatic metrics

Development cycle

1. estimate models independently on training data
2. adjust λ_k on tuning data by optimizing on quality

Minimum error rate training

$$s(H,R) = \sum_k \lambda_k F_k(H,R,\mathbf{a})$$

1. compute n-best hypotheses with fixed λ: $\{H_1 \ldots H_n\}$ s.t. $s(H_1,\lambda) > s(H_2,\lambda) > ... > s(H_n,\lambda)$
2. evaluation these n hypotheses $q(H_1,R) \ldots q(H_n,R)$
3. adjust λ s.t. $q(H_i,R) > q(H_j,R) \Rightarrow s(H_i,\lambda) > s(H_j,\lambda)$
4. convergence ? or back to 1

$\Rightarrow$ difficult optimization problem

$\Rightarrow$ requires suitable automatic metrics

Evaluation for MT: an active field

Market analysis

- a clear leader: BLEU (since 2001)
- good contenders: METEOR (2004), TER (2005)
- many others: GTM, DDM, BLANC, PER, MT-NCD, ATEC, HTER TESLA, ULC, TERP, SEPIA, IQTM, BEWT-E, LRK4, MEANT, etc.

General principle

Compare a translation candidate (H) with one or several human (reference) translations (R): $\Rightarrow q(R, H)$

Evaluation for MT: an active field

Market analysis

- a clear leader: BLEU (since 2001)
- good contenders: METEOR (2004), TER (2005)
- many others: GTM, DDM, BLANC, PER, MT-NCD, ATEC, HTER TESLA, ULC, TERP, SEPIA, IQTM, BEWT-E, LRK4, MEANT, etc.

General principle

Compare a translation candidate (H) with one or several human (reference) translations (R): $\Rightarrow q(R, H)$

BLEU: example

Reference translations

1. It is a guide to action that ensures that the military will forever heed Party commands.
2. It is the guiding principle which guarantees the military forces always being under the command of the Party.
3. It is the practical guide for the army always to heed the directions of the party

H_1 : It is to insure the troops forever hearing the activity guidebook that party direct.
H_2 : It is a guide to action which ensures that the military always obeys the command of the party.

BLEU: example

Reference translations

1. It is a guide to action that ensures that the military will forever heed Party commands.
2. It is the guiding principle which guarantees the military forces always being under the command of the Party.
3. It is the practical guide for the army always to heed the directions of the party

H_1 : It is to insure the troops forever hearing the activity guidebook that party direct.
H_2 : It is a guide to action which ensures that the military always obeys the command of the party.

BLEU: example

Reference translations

1. It is a guide to action that ensures that the military will forever heed Party commands.
2. It is the guiding principle which guarantees the military forces always being under the command of the Party.
3. It is the practical guide for the army always to heed the directions of the party

H_1 : It is to insure the troops forever hearing the activity guidebook that party direct.

H_2 : It is a guide to action which ensures that the military always obeys the command of the party.

BLEU: example

H_1 : It is to insure the troops forever hearing the activity guidebook that party direct.

H_2 It is a guide to action which ensures that the military always obeys the command of the party.

Conclusion: $H_2 \gg H_1$

BLEU: definition

- modified n-gram precision

$$\exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

- brevity penalty (BP)

$$\mathrm{BP} = \begin{cases} 1 & \text{if hypothesis is longer than reference translation} \\ e^{1-\frac{r}{c}} & \text{otherwise} \end{cases}$$

- score calculation

$$\textsc{Bleu} = \mathrm{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Like it or not, you have to use it [4]

[4] John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. (2004) Confidence Estimation for Machine Translation. In Proceedings of Coling 2004, Geneva, August 2004, pp. 315–321

BLEU: definition

- modified n-gram precision

$$\exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

- brevity penalty (BP)

$$\text{BP} = \begin{cases} 1 & \text{if hypothesis is longer than reference translation} \\ e^{1-\frac{r}{c}} & \text{otherwise} \end{cases}$$

- score calculation

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Like it or not, you have to use it [4]

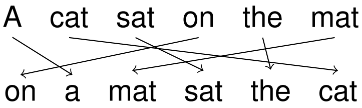
[4] John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. (2004) Confidence Estimation for Machine Translation. In Proceedings of Coling 2004, Geneva, August 2004, pp. 315–321

BLEU: discussion

- scores are not straightforward to interpret
  - BLEU = 30 ... so what ?
  - depend on many factors: number of reference translations, language, tokenization, etc.
- favor statistical systems
- syntax poorly modeled
- comparison at the string level
- cannot be decomposed at the sentence level

METEOR: an IR-inspired metric

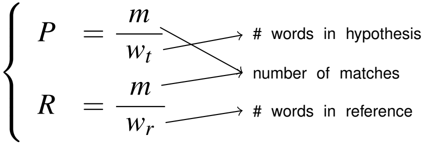
1. word-to-word alignment between hypothesis and reference



A cat sat on the mat

on a mat sat the cat

2. evaluation by n-gram recall/precision



$$\begin{cases} P = \dfrac{m}{w_t} \\ R = \dfrac{m}{w_r} \end{cases}$$

# words in hypothesis

number of matches

# words in reference

3. using several reference translations: max over all references

METEOR: computing alignments

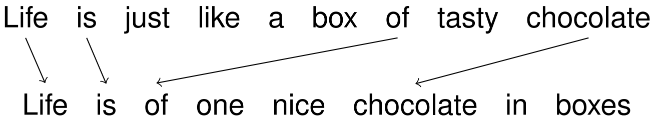
- rule-based approach
- seach for $1 \leftrightarrow 1$ associations
- allow flexible associations :
  1. identical words in hypothesis and reference
  2. words from the same morphological family
  3. synomyms

Life is just like a box of tasty chocolate

Life is of one nice chocolate in boxes

METEOR: computing alignments

- rule-based approach
- seach for $1 \leftrightarrow 1$ associations
- allow flexible associations :
  1. identical words in hypothesis and reference
  2. words from the same morphological family
  3. synomyms

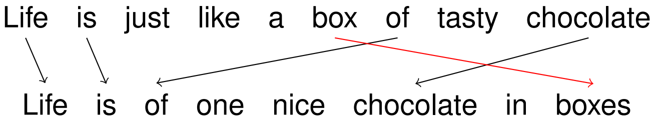


Life is just like a box of tasty chocolate

Life is of one nice chocolate in boxes

METEOR: computing alignments

- rule-based approach
- seach for $1 \leftrightarrow 1$ associations
- allow flexible associations :
  1. identical words in hypothesis and reference
  2. words from the same morphological family
  3. synomyms

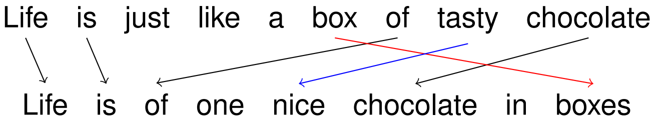


Life is just like a box of tasty chocolate

Life is of one nice chocolate in boxes

METEOR: computing alignments

- rule-based approach
- seach for $1 \leftrightarrow 1$ associations
- allow flexible associations :
  1. identical words in hypothesis and reference
  2. words from the same morphological family
  3. synomyms



Life is just like a box of tasty chocolate

Life is of one nice chocolate in boxes

METEOR: assessing grammaticality

- alignments are independent from position
  $\Rightarrow$ all permutations of an hypothesis have the same score
- penalize discontinuous matches
  - segmentation in phrases
  - measure of fragmentation

$$\text{frag} = \frac{\text{\# of phrases}}{\text{\# of pairs}}$$

  - final score final

$$\text{score} = (1 - \gamma \cdot \text{frag}^{\beta}) \times \underbrace{\frac{P \times R}{\alpha \cdot P + (1 - \alpha) \cdot R}}_{F_{\text{mean}}}$$

- 3 parameters: α, β, γ

METEOR : discussion

- better correlation with human judgments
- phrase-level score
- tunable for each language pair
- still difficult to interpret
- relies on linguistic resources

TER: Translation Edit Rate

Simulate post-editing

- H : translation hypothesis
- R : reference translation
- TER: minimal number of edits to transform H into R
  - word deletion
  - word substitution
  - word insertion
  - bloc movement
- generalization of Word Error Rate

$$TER = \frac{\text{\# of edits}}{\text{average \# of reference words}}$$

TER: example

H :
To bring an end to military conflict on October 6 on a a comprehensive blockade against Palestine

R :
To bring an end to military conflict, the Israeli military began a comprehensive blockade against Palestine on October 6.

TER: example

H :

To bring an end to military conflict on October 6 on a a comprehensive blockade against Palestine

R :

To bring an end to military conflict , the Israeli military began a comprehensive blockade against Palestine on October 6 .

insertion	4
substitution	1
deletion	1
block movement	1

TER: discussion

- evaluation close to a real task (post-editing)
- results are more interpretable than for other metrics
- insensitive to semantic closeness
- complexity of computation ⇒ approximate search

Extensions: TERP, HTER

HTER: Human Translation Edit Rate

Context

- TER heavily depends on the reference translation
- perform human post-editing to transform the output of a system into the closest acceptable translation
- HTER measures TER between the original hypothesis and the new reference translation
- this can be applied to most metrics (e.g. hBLEU, hMETEOR)

HTER: human post-edition

Possible conditions

- possibly several post-editors
- trained to produce as few corrections as possible
- may not have access to the source sentence
- post-editions can be post-edited anew (without knowledge of any other reference translation)

HTER: discussion

- does not rely on simple string matching (*human in the loop*)
- can be interpreted
- depends on the targeted application (defines acceptable quality)
- humans may know the target language only
- but costly (and difficult), and results are non-reproducible

MT evaluation: current directions

Machine-learned metrics

- multi-criteria evaluation $\Rightarrow \{G_k(H,R), k=1\dots K\}$
- find $\lambda_1 \dots \lambda_k$ s.t. $\sum_k \lambda_k G_k(H,R)$ correlates well with human judgments

Evaluation of evaluation metrics

- MetricsMaTR (2009; 2010); WMT (2010; 2011)
- correlation with human judgments or rankings

MT evaluation: current directions

Machine-learned metrics

- multi-criteria evaluation $\Rightarrow \{G_k(H,R), k = 1 \ldots K\}$
- find $\lambda_1 \ldots \lambda_k$ s.t. $\sum_k \lambda_k G_k(H,R)$ correlates well with human judgments

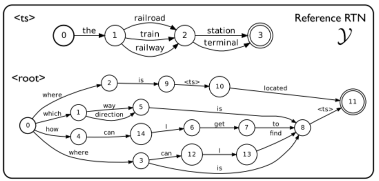
Evaluation of evaluation metrics

- MetricsMaTR (2009; 2010); WMT (2010; 2011)
- correlation with human judgments or rankings

MT evaluation: current directions

Robust evaluation

- use automatic synonyms or paraphrases (METEOR, TERP) for evaluation and/or tuning
- pre-encode huge numbers of reference translations into lattices (HYTER: meaning-equivalent networks)



- similarities with multi-source translation (e.g. acquire source paraphrases from monolingual contributors)

MT evaluation: current directions

Error analysis

- automatic error detection
- automatic error classification
  e.g. missing word, word order, incorrect word (sense, incorrect form, style, etc.), unknown word

Confidence estimation

- system-, sentence-, n-gram-, word-level

MT evaluation: current directions

Error analysis

- automatic error detection
- automatic error classification
  e.g. missing word, word order, incorrect word (sense, incorrect form, style, etc.), unknown word

Confidence estimation

- system-, sentence-, n-gram-, word-level

Conclusion

Automatic metrics

- necessary for developing systems
- very active field of research
- much progress ahead

Human judgments

- necessary to evaluation the true performance of systems
- ... and to develop automatic metrics

Conclusion

Automatic metrics

- necessary for developing systems
- very active field of research
- much progress ahead

Human judgments

- necessary to evaluation the true performance of systems
- ... and to develop automatic metrics