A Brief Introduction to Machine Translation Evaluation

Francisco Guzmán

ALT Research Group

Qatar Computing Research Institute (QCRI)

MT Marathon of the Americas

Urbana-Champaign, IL, USA

May 12, 2015

Special Thanks...

to **Cristina España i Bonet** and **Lluís Màrquez**
for some of the slides

to the **QCRI-ALT** group
for their feedback

What to Expect Today

- Why is evaluating MT a hard task?
- How do we (humans) evaluate translations?
- What are different approaches for automatic MT eval?
- What are (dis-)advantages of automatic MT eval?

Can You Evaluate This Translation?

Source:
Renzi logra una nueva ley electoral para dar estabilidad a Italia

Candidate/Hypothesis:
Renzi achieved a new electoral law to give stability to Italy

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

What Makes a Good Translation?

According to **professional translators**, it all depends...

- guidelines (i.e. client requirements)
- genre (e.g. news, blog)
- style (e.g. humorous, wordy, scientific)
- localization (e.g. tailored for target audience)
- ...

Not an easy task!

Difficulties of MT Evaluation

- Machine Translation is an **open** NLP task
  - the *correct translation* is not unique
  - the set of admissible translations can be large
  - translation correctness is not black and white

Difficulties of MT Evaluation

- Machine Translation is an **open** NLP task
  - the *correct translation* is not unique
  - the set of admissible translations can be large
  - translation correctness is not black and white
- Evaluation is necessary in the MT system development cycle

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

What Makes a Good Automatic Translation?

Idea: Compare MT output to a human reference

Source:
Renzi logra una nueva ley electoral para dar estabilidad a Italia

Candidate/Hypothesis:
Renzi achieved a new electoral law to give stability to Italy

Reference:
Renzi passed new electoral law aimed to stabilize Italy

What Makes a Good Automatic Translation?

Idea: Compare MT output to a human reference

Source:
Renzi logra una nueva ley electoral para dar estabilidad a Italia

Candidate/Hypothesis:
Renzi achieved a new electoral law to give stability to Italy

Reference:
Renzi passed new electoral law aimed to stabilize Italy

This is a simpler task

MT Evaluation

Setting Compute **similarity** between system's output and one or several reference translations

Challenge The similarity measure should be able to discriminate whether the two sentences convey the same meaning

MT Evaluation

Setting Compute **similarity** between system's output and one or several reference translations

Challenge The similarity measure should be able to discriminate whether the two sentences convey the same meaning

two possibilities: **manual** and **automatic evaluation**

Talk Overview

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Different Views on Quality

Adequacy (or Fidelity) Does the output convey the same meaning as the input sentence? Is part of the message lost, added, or distorted?

Fluency (or Intelligibility) Is the output fluent? This involves both grammatical correctness and idiomatic word choices.

Post–edition effort Time required to *repair* the translation, number of key strokes, etc.

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو مؤسسة قطر

Manual Evaluation: TAUS recommendation

Adequacy How much of the meaning expressed in the gold-standard translation or the source is also expressed in the target translation?

4	Everything
3	Most
2	Little
1	None

Fluency To what extent is a target side translation grammatically well informed, without spelling errors and experienced as using natural/intuitive language by a native speaker?

4	Flawless
3	Good
2	Disfluent
1	Incomprehensible

Other examples: NIST

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Ranking

Pairwise

Annotators chose the best system, given the source and target sentence, and 2 anonymised random systems.

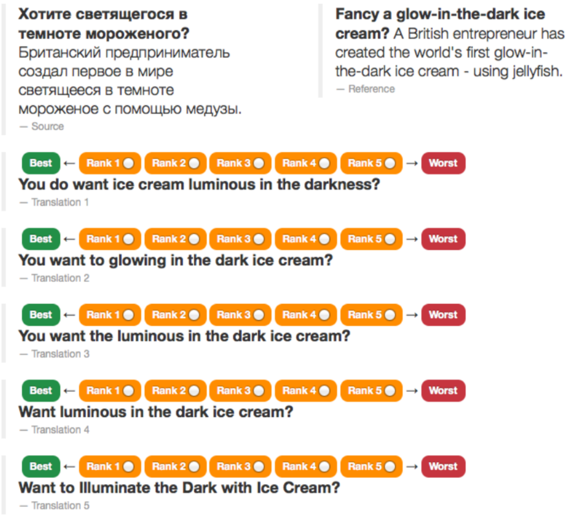
N-way

Annotators rank *n* anonymised systems, randomly selected and randomly ordered.

Ranking with Appraise

(Federmann,2012)

Appraise



Ranking is better

Advantages:

- Conceptually easier to rank
- Higher agreement among annotators (Callison-Burch et al., 2007)
- No scales to be defined

Disadvantages:

- Less information is provided

Manual Evaluation
HTER

Human-targeted Translation Error Rate, HTER

Annotation Post-edition of the candidate translation to have the same meaning as a reference translation with as few edits as possible

Evaluation TER with the candidate translation and the post-edited reference

$$HTER = \frac{\text{Substitutions} + \text{Insertions} + \text{Deletions} + \text{Shifts}}{\text{ReferenceWords}}$$

Evaluation matters!

Progress in the field is measured by evaluation campaigns:

NIST Open Machine Translation Evaluation

WMT Workshop Machine Translation

IWSLT International Workshop on Spoken Language Translation

Human Evaluation Shortcomings

- Subjective
- Costly
- Non-reusable

Human Evaluation Shortcomings

- Subjective
- Costly
- Non-reusable

Talk Overview

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو مؤسسة قطر

Reference-based Automatic Evaluation (RAE)

Setting

⇒ Compute similarity between MT **system's output** (Hyp) and one or several **reference** translations (Ref)

Source	`Es un plan de acción que asegura que el Ejército siempre cumpla las órdenes del partido`
Hypothesis	`It is a guide to action which ensures that the military always obeys the commands of the party.`
Reference 1	`It is a guide to action that ensures that the military will forever heed Party commands .`
Reference 2	`It is the guiding principle which guarantees the military forces always being under the command of the Party.`

Reference-based Automatic Evaluation (RAE)

Setting

⇒ Compute similarity between MT **system's output** (Hyp) and one or several **reference** translations (Ref)

Source	`Es un plan de acción que asegura que el Ejército siempre cumpla las órdenes del partido`
Hypothesis	`It is a guide to action which ensures that the military always obeys the commands of the party.`
Reference 1	`It is a guide to action that ensures that the military will forever heed Party commands .`
Reference 2	`It is the guiding principle which guarantees the military forces always being under the command of the Party.`

Challenge

⇒ The similarity measure should be able to discriminate whether the two sentences convey the same meaning

Desiderata for MT Metrics
(Lavie, 2009)

Human-like	High-levels of correlation with quantified human notions of translation quality
Fine-grained	Sensitivity to small differences in MT quality between systems and versions of systems
Consistency	Same MT system on similar texts should produce similar scores
Reliability	MT systems that score similarly will perform similarly
Lightweight	Fast, easy to run

Desiderata for MT Metrics
(Lavie, 2009)

Human-like	High-levels of correlation with quantified human notions of translation quality
Fine-grained	Sensitivity to small differences in MT quality between systems and versions of systems
Consistency	Same MT system on similar texts should produce similar scores
Reliability	MT systems that score similarly will perform similarly
Lightweight	Fast, easy to run

Different Levels of Analysis

- Lexical (words)
- Syntactic
- Semantic
- Pragmatic (discourse)

Lexical Matching

First approaches

⇒ **Lexical similarity** as a measure of quality

- word n-gram matching, edit distance, etc.
- **BLEU**, NIST, TER, Meteor, Rouge, etc.
- (Papineni et al., 2002; Doddington, 2002; Snover et al., 2006; Lavie & Agarwal 2007; Lin, 2004; etc.)

Lexical Matching

First approaches

⇒ **Lexical similarity** as a measure of quality

- word n-gram matching, edit distance, etc.
- **BLEU**, NIST, TER, Meteor, Rouge, etc.
- (Papineni et al., 2002; Doddington, 2002; Snover et al., 2006; Lavie & Agarwal 2007; Lin, 2004; etc.)

Nowadays, BLEU is accepted as *the de-facto* standard metric.

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

IBM BLEU

BLEU: a Method for Automatic Evaluation of Machine Translation

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu
IBM Research Division

"The main idea is to use a weighted average of variable length phrase matches against the reference translations. This view gives rise to a family of metrics using various weighting schemes. We have selected a promising baseline metric from this family."

Automatic evaluation

IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

BiLingual Evaluation Understudy, BLEU

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log \mathrm{P}_n\right)$$

- Precision at different levels (n=1: unigrams, n=2: bigrams, etc)
- Geometric average of P_n (empirical suggestion)
- w_n positive weights summing to one (typically $1/N$)
- Brevity penalty

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

IBM BLEU

Hypothesis:

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

Reference 1:

`It is a guide to action that ensures that the military will forever heed Party commands.`

Reference 2:

`It is the guiding principle which guarantees the military forces always being under the command of the Party.`

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

IBM BLEU

Hypothesis:

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

Reference 1:

It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2:

It is the guiding principle which guarantees the military forces always being under the command of the Party.

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

IBM BLEU

Modified n-gram precision (1-gram)

Precision-based measure, but:

Candidate:
`The the the the the the the.`

Reference 1:
`The cat is on the mat.`

Reference 2:
`There is a cat on the mat.`

IBM BLEU

Modified n-gram precision (1-gram)

Precision-based measure, but: $\text{Prec.} = \frac{1+}{7}$

Candidate:

`The the the the the the the.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

IBM BLEU

Modified n-gram precision (1-gram)

Precision-based measure, but: Prec. $= \frac{2 +}{7}$

Candidate:

`The the the the the the the.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

IBM BLEU

Modified n-gram precision (1-gram)

Precision-based measure, but:

$$\text{Prec.} = \frac{3 +}{7}$$

Candidate:

`The the the the the the the.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

IBM BLEU

Modified n-gram precision (1-gram)

Precision-based measure, but: $\text{Prec.} = \frac{4+}{7}$

Candidate:

`The the the the the the the.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

Qatar Computing Research Institute

Member of Qatar Foundation

IBM BLEU

Modified n-gram precision (1-gram)

Precision-based measure, but: $\text{Prec.} = \frac{5+}{7}$

Candidate:

`The the the the the the the.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

Qatar Computing Research Institute
Member of Qatar Foundation

IBM BLEU

Modified n-gram precision (1-gram)

Precision-based measure, but: $\text{Prec.} = \frac{6+}{7}$

Candidate:

`The the the the the the the.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

IBM BLEU

Modified n-gram precision (1-gram)

Precision-based measure, but: $\text{Prec.} = \frac{7}{7}$

Candidate:

`The the the the the the the.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو مؤسسة قطر

IBM BLEU

Modified n-gram precision (1-gram)

A reference word should only be matched once.

Algorithm:

1. Count number of times w_i occurs in the candidate.
2. Keep the minimum of (1) and the maximum number of times w_i appears in any reference (*clipping*).
3. Add these values and divide by candidate's number of words.

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

IBM BLEU

Modified n-gram precision (1-gram)

Modified 1-gram precision:

Candidate:

`The the the the the the the`

Reference 1:

`The cat is on the mat`

Reference 2:

`There is a cat on the mat`

1. $w_i \rightarrow \text{The}$
   $\#w_{i,R1} = 2$
   $\#w_{i,R2} = 1$
   $\#w_{i,C} = 7$
2. $\text{Max}_{(R^*)}=2$,
   $\Rightarrow \text{Min}(R^*,c)=2$
3. No more distinct words

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

IBM BLEU

Modified n-gram precision (1-gram)

Modified 1-gram precision: $P_1 =$

Candidate:

`The the the the the the the`

Reference 1:

`The cat is on the mat`

Reference 2:

`There is a cat on the mat`

1. $w_i \rightarrow$ The
   $\#w_{i},R1 = 2$
   $\#w_{i},R2 = 1$
   $\#w_{i},C = 7$
2. $\text{Max}_{(R^*)}=2$,
   $\Rightarrow \text{Min}(R^*,c)=2$
3. No more distinct words

Qatar Computing Research Institute

Member of Qatar Foundation

IBM BLEU

Modified n-gram precision (1-gram)

Modified 1-gram precision: $P_1 = \frac{2}{}$

Candidate:

`The the the the the the the`

Reference 1:

`The cat is on the mat`

Reference 2:

`There is a cat on the mat`

1. $w_i \rightarrow$ The
   $\#w_i,R1 = 2$
   $\#w_i,R2 = 1$
   $\#w_i,C = 7$
2. $\text{Max}_{(R^*)}$=2,
   $\Rightarrow \text{Min}(R^*,c)$=2
3. No more distinct words

Qatar Computing Research Institute
معهد قطر لبحوث الحوسبة
Member of Qatar Foundation عضو في مؤسسة قطر

IBM BLEU

Modified n-gram precision (1-gram)

Modified 1-gram precision: $\mathrm{P}_1 = \frac{2}{7}$

Candidate:

`The the the the the the the`

Reference 1:

`The cat is on the mat`

Reference 2:

`There is a cat on the mat`

1. $w_i \rightarrow$ The
   $\#w_i, R1 = 2$
   $\#w_i, R2 = 1$
   $\#w_i, C = 7$
2. $\text{Max}_{(R^*)}=2$,
   $\Rightarrow \text{Min}(R^*,\text{c})=2$
3. No more distinct words

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Extending to n-grams

- Generalisation to multiple sentences:

$$\mathrm{P}_n = \frac{\sum_{C \in \{\text{candidates}\}} \sum_{ngram \in C} Count_{\text{clipped}}(ngram)}{\sum_{C \in \{\text{candidates}\}} \sum_{ngram \in C} Count(ngram)}$$

low n	**high** n
adequacy	fluency

Automatic evaluation

IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

Brevity penalty

Candidate:

`of the`

Reference 1:

`It is a guide to action that ensures that the military will forever heed Party commands`

Reference 2:

`It is the guiding principle which guarantees the military forces always being under the command of the Party`

Reference 3:

`It is the practical guide for the army always to heed the directions of the party`

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Automatic evaluation

IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

Brevity penalty

Candidate:

`of the`

$P_1 = 2/2, P_2 = 1/1$

Reference 1:

`It is a guide to action that ensures that the military will forever heed Party commands`

Reference 2:

`It is the guiding principle which guarantees the military forces always being under the command of the Party`

Reference 3:

`It is the practical guide for the army always to heed the directions of the party`

Automatic evaluation

IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

Brevity penalty

$$\mathrm{BP} = \begin{cases} 1 & \text{if} \quad c > r \\ e^{1-r/c} & \text{if} \quad c \leq r \end{cases}$$

c candidate length, r reference length

- Multiplicative factor
- At sentence level, huge punishment for short sentences
- Estimated at document level

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Sentence-level BLEU

Sometimes we want to evaluate BLEU at the sentence level
This can lead to trouble:

- **Problem**
  - Precision: Zero matches = Zero score
- **Solution**
  - Smooth Precision : Add + 1 to precision counts
  - Smooth BP : Add +1 to reference component

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Limits of lexical similarity

Hyp: This sentence is going to be difficult to evaluate.

Ref1: The evaluation of the clause is complicated.
Ref2: The sentence will be hard to qualify.
Ref3: The translation is going to be hard to evaluate.
Ref4: It will be difficult to punctuate the output.

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Limits of lexical similarity

Hyp: `This sentence is going to be difficult to evaluate.`

Ref1: `The evaluation of the clause is complicated.`
Ref2: `The sentence will be hard to qualify.`
Ref3: `The translation is going to be hard to evaluate.`
Ref4: `It will be difficult to punctuate the output.`

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Limits of lexical similarity

Hyp: This sentence is going to be difficult to evaluate.

Ref1: The evaluation of the clause is complicated.
Ref2: The sentence will be hard to qualify.
Ref3: The translation is going to be hard to evaluate.
Ref4: It will be difficult to punctuate the output.

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Metric for Evaluation of Translation with Explicit ORdering

$$METEOR = (1 - Pen)F_\alpha$$

$$F_\alpha = \frac{PR}{\alpha P + (1-\alpha)R}$$

*P*recision and *R*ecall weighted harmonic mean

$$Pen = \gamma \left(\frac{\text{chunks}}{\text{mapped unigrams}} \right)^\beta$$

Penalty factor, penalises non-contiguous matches

Matches: exact, lemma, synonym, paraphrase

Metric for Evaluation of Translation with Explicit ORdering

$$\mathit{METEOR} = (1 - \mathit{Pen})F_\alpha$$

$$F_\alpha = \frac{PR}{\alpha P + (1-\alpha)R}$$

*P***recision** and *R***ecall** weighted harmonic mean

$$\mathit{Pen} = \gamma \left(\frac{\text{chunks}}{\text{mapped unigrams}} \right)^\beta$$

Penalty factor, penalises non-contiguous matches

Matches: exact, lemma, synonym, paraphrase

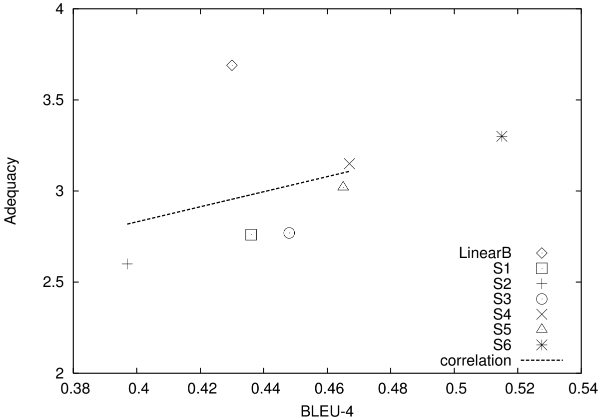
Problems of Lexical Similarity Measures

- Lexical similarity is nor a **sufficient** neither a **necessary** condition so that two sentences express the same meaning (Culy and Riehemann, 2003; Coughlin, 2003; Callison-Burch et al., 2006)
- The **reliability** of lexical metrics depends very strongly on the **heterogeneity/representativity** of reference translations
- Lexical metrics have problems distinguishing MT output from fully fluent and adequate translations obtained from them through professional postediting (Denkowski and Lavie, 2012)

Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise

(Callison-Burch et al., 2006; Koehn and Monz, 2006)



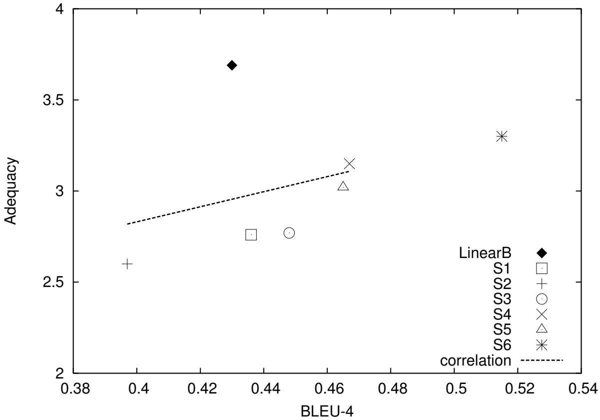
Qatar Computing Research Institute

Member of Qatar Foundation

Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise

(Callison-Burch et al., 2006; Koehn and Monz, 2006)



Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise

(Callison-Burch et al., 2006; Koehn and Monz, 2006)

⇒ *n*-gram based metrics favor MT systems which closely replicate the lexical realization of the references

⇒ Test sets tend to be similar (domain, register, sublanguage) to training materials

⇒ Statistical MT systems heavily rely on the training data

⇒ Statistical MT systems tend to share the reference sublanguage and be favored by *n*-gram based measures

Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise
(Callison-Burch et al., 2006; Koehn and Monz, 2006)

⇒ *n*-gram based metrics favor MT systems which closely replicate the lexical realization of the references

⇒ Test sets tend to be similar (domain, register, sublanguage) to training materials

⇒ Statistical MT systems heavily rely on the training data

⇒ Statistical MT systems tend to share the reference sublanguage and be favored by *n*-gram based measures

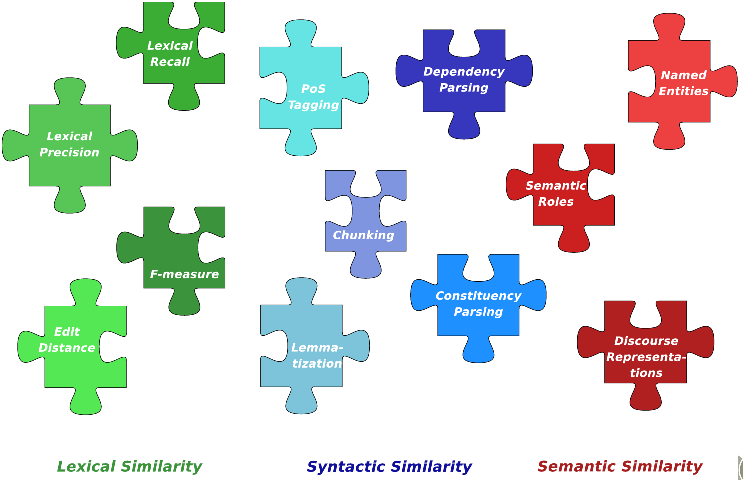
Linguistic Generalization

Active area of research

⇒ Generalization over lexical matching and usage of more complex linguistic information to compute similarity

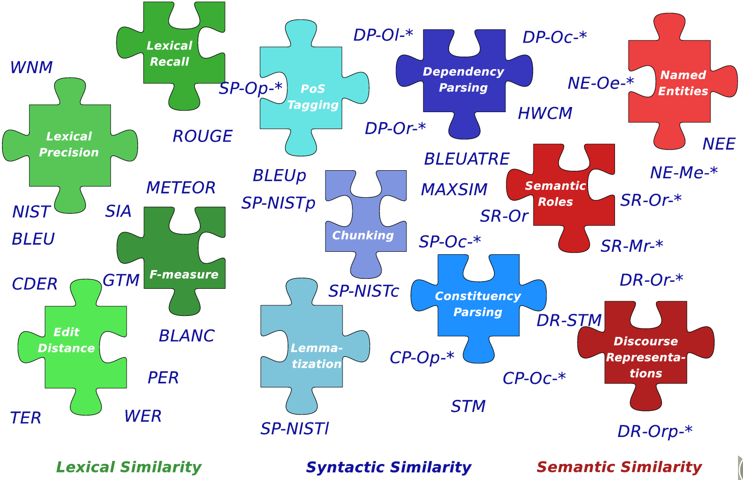
- stemming, synonymy, paraphrasing, etc.
- shallow parsing, constituency and dependency parsing, named entities, semantic roles, textual entailment, etc.
- discourse trees

Existing Metrics



معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Existing Metrics



معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Talk Overview

Which one is better?

Idea: Measure the *correlation* of evaluation metrics with human judgments (e.g. Appraise)

Campaigns:

- metricsMATR (NIST)
- WMT metrics

	Metric	Orig.
II	SEMPOS	.902
	AMBER	.857
	METEOR	.834
	TERRORCAT	.831
	SIMPBLEU	.823
	TER	.812
	BLEU	.810
	POSF	.754
	...	
III	NIST	.817
	...	
IV	Asiya-LEX	.879
	...	

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو مؤسسة قطر

Which one is better?

Idea: Measure the *correlation* of evaluation metrics with human judgments (e.g. Appraise)

Campaigns:

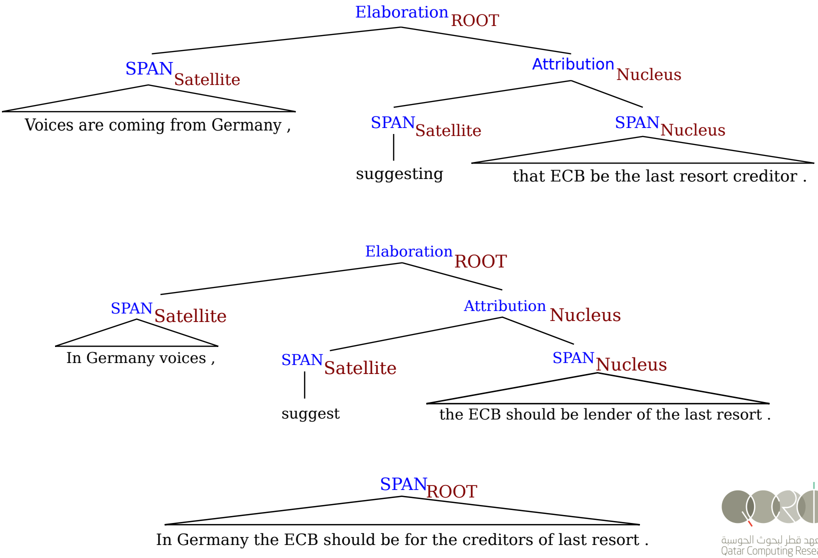
- metricsMATR (NIST)
- WMT metrics

	Metric	Orig.
II	SEMPOS	.902
	AMBER	.857
	METEOR	.834
	TERRORCAT	.831
	SIMPBLEU	.823
	TER	.812
	BLEU	.810
	POSF	.754
	...	
III	NIST	.817
	...	
IV	Asiya-LEX	.879
	...	

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Going upwards: Discourse

Guzmán et al, ACL2014



معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Setting

- Discourse structures: computed at sentence level with the RST-based parser from Joty et al. (2012)
- Similarity: computed with STK kernel (Collins & Duffy, 2001)
  ⇒ the similarity is the sum of all common sub-trees

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Setting

- Discourse structures: computed at sentence level with the RST-based parser from Joty et al. (2012)
- Similarity: computed with STK kernel (Collins & Duffy, 2001)
  $\Rightarrow$ the similarity is the sum of all common sub-trees

Untuned combinations

[WMT12, into-*en*, system-level, ρ]

- Combination with other existing evaluation metrics
- Other *smarter* ways are possible.

	Metric	Orig.	+DR-LEX
I	DR-LEX	.876	–
II	SEMPOS	.902	**.903**
	AMBER	.857	**.869**
	METEOR	.834	**.888**
	TERRORCAT	.831	**.889**
	SIMPBLEU	.823	**.859**
	TER	.812	**.848**
	BLEU	.810	**.846**
	POSF	.754	**.857**
	...		
III	NIST	.817	**.875**
	...		
IV	Asiya-LEX	.879	**.882**
	...		
	average	**.839**	**.874**
	diff.		**+.035**

Talk Overview

1. Motivation
2. Manual Evaluation
3. Automatic Evaluation
4. Recent advances
5. Conclusions
6. Extra slides

Virtues and curse

⇒ Automatic evaluation metrics have notably accelerated the development cycle of MT systems
- ▷ Cheap, objective and reusable
- ▷ Used for error analysis, system optimization, system comparison, etc.

⇒ Risks of Automatic Evaluation
- ▷ System over-tuning
- ▷ Blind system development
- ▷ Unfair system comparisons

MT Evaluation

Summary

- Evaluation is important in the system development cycle. Automatic evaluation accelerates significantly the process.
- Manual evaluation is still necessary but shows low agreements among annotators
- Up to now, most (common) metrics rely on lexical similarity, but it cannot assure a correct evaluation.
- Current work is being devoted to go beyond lexical similarity.

Thank you!

A Brief Introduction to Machine Translation Evaluation

Francisco Guzmán
ALT Research Group
Qatar Computing Research Institute (QCRI)

MT Marathon of the Americas
Urbana-Champaign, IL, USA
May 12, 2015

Learning with structured/distributed representations

Goal Instead of adjusting weights of already existing metrics, we want to work in a unified learning framework, able to represent many layers of linguistic information and able to learn from fine-grained features

- Two alternatives for the input representation
  - ⇒ *Structured* (with kernel-based learning)
  - ⇒ *Distributed* (with ANN learning)
- Common setting: pairwise quality comparison

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Differentiating *better* from *worse* translation

- Input: $\langle t_1, t_2, r \rangle$
  - ⇒ "Is t_1 a better translation than t_2, given r"?
- Pairwise ranking setting
  - ⇒ closer to the evaluation that humans do better
  - ⇒ valid for most MT comparison/ranking tasks
  - ⇒ not an absolute quality score

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Learning with preference kernels

Guzmán et al, EMNLP2014

- Tree-based representation of all layers of information
- Pairwise ranking with the preference kernel (Shen & Joshi, 2003)
- Learning example: $\langle h_1, h_2 \rangle = \langle \phi_M(t_1, r), \phi_M(t_2, r) \rangle$

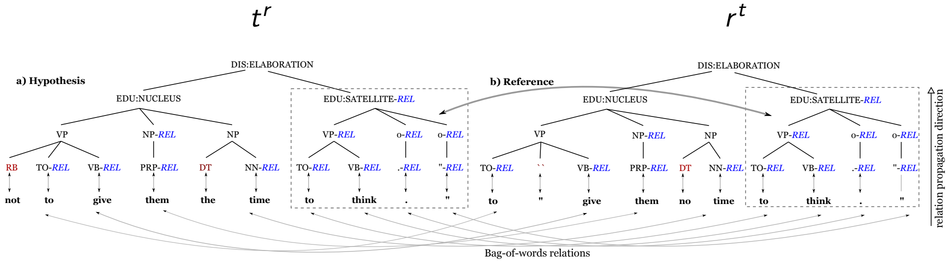
  $\Rightarrow$ ϕ_M makes a structured and relational representation of t and r

  $\Rightarrow$ $\phi_M(t_1, r) = \langle t_1^r, r^{t_1} \rangle$

  $\Rightarrow$ two separate trees instead of a graph

Learning with preference kernels: $\phi_M(t, r)$

Guzmán et al, EMNLP2014



Learning with preference kernels (II)

Guzmán et al, EMNLP2014

- Learning example: $\langle h_1, h_2 \rangle = \langle \phi_M(t_1, r), \phi_M(t_2, r) \rangle$
- Preference kernel (Shen & Joshi, 2003)
  - $PK(\langle h_1, h_2 \rangle, \langle h_1', h_2' \rangle) =$
    $$K(h_1, h_1') + K(h_2, h_2') - K(h_1, h_2') - K(h_2, h_1')$$
  - $K(h_1, h_1') = PTK(t_1^r, t_1'^r) + PTK(r^{t_1}, r^{t_1'})$
  - PTK = Partial Tree Kernel (Moschitti, 2006)

Qatar Computing Research Institute
Member of Qatar Foundation

Learning with preference kernels (II)

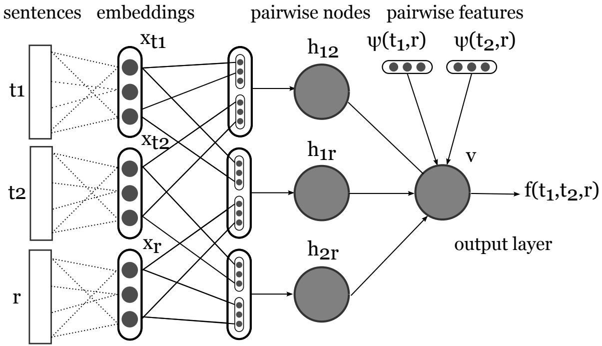
Guzmán et al, EMNLP2014

- Learning example: $\langle h_1, h_2 \rangle = \langle \phi_M(t_1, r), \phi_M(t_2, r) \rangle$
- Preference kernel (Shen & Joshi, 2003)
  - $PK(\langle h_1, h_2 \rangle, \langle h'_1, h'_2 \rangle) =$
    $$K(h_1, h'_1) + K(h_2, h'_2) - K(h_1, h'_2) - K(h_2, h'_1)$$
  - $K(h_1, h'_1) = PTK(t_1^r, t_1'^r) + PTK(r^{t_1}, r^{t'_1})$
  - PTK = Partial Tree Kernel (Moschitti, 2006)

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Learning with distributed representations and NNs

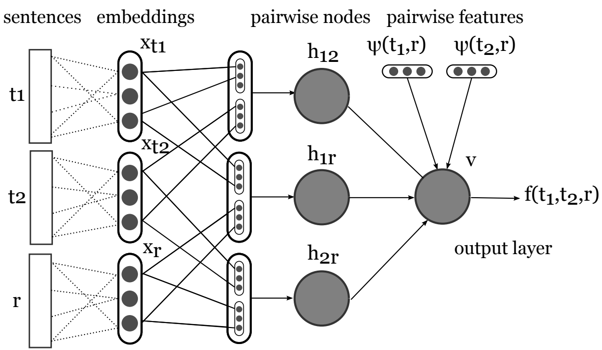
Guzmán et al, ACL2015



- Input mapped to fixed-length vectors $[\mathbf{x}_{t1}, \mathbf{x}_{t2}, \mathbf{x}_r]$ using syntactic (Stanford's parser) and semantic embeddings (a la *'word2vec'*)

Learning with distributed representations and NNs

Guzmán et al, ACL2015

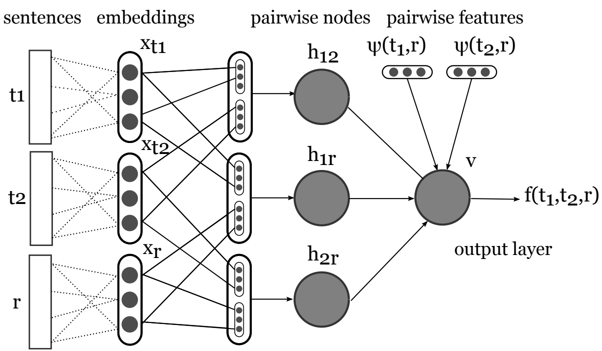


- Hidden layer to compute three types of interactions: $\mathrm{sim}(t_1, r)$, $\mathrm{sim}(t_2, r)$, and $\mathrm{sim}(t_1, t_2)$.

Qatar Computing Research Institute
Member of Qatar Foundation

Learning with distributed representations and NNs

Guzmán et al, ACL2015



- External sources of information as direct features (*skip arcs*). We plug in BLEU, NIST, TER, and METEOR scores.