Deep evaluation of hybrid architectures: Use of different metrics in MERT weight optimization

Cristina España-Bonet, Gorka Labaka,
Arantza Díaz de Ilarraza, Kepa Sarasola, Lluís Màrquez

Free/Open-source Rule-based Machine Translation

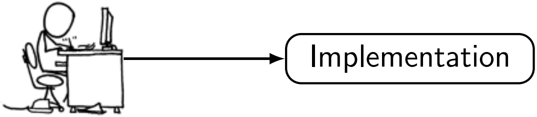
Gothenburg, June 14th, 2012

Overview

1. Motivation
2. SMatxinT, a hybrid translator
3. Systems' evaluation
4. Conclusions

Motivation

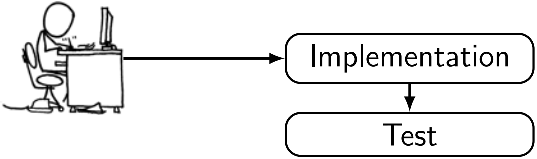
Evaluation during system development



Implementation

Motivation

Evaluation during system development

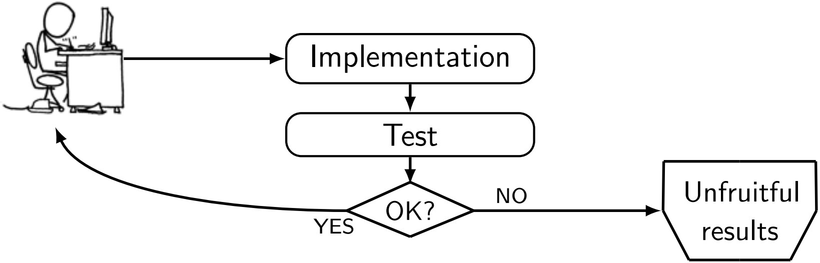


Implementation

Test

Motivation

Evaluation during system development



Implementation

Test

OK?

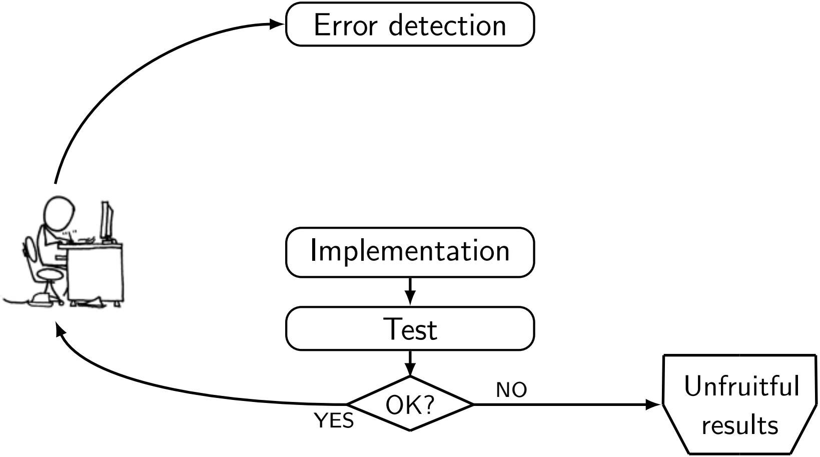
NO

YES

Unfruitful results

Motivation

Evaluation during system development



Error detection

Implementation

Test

OK?

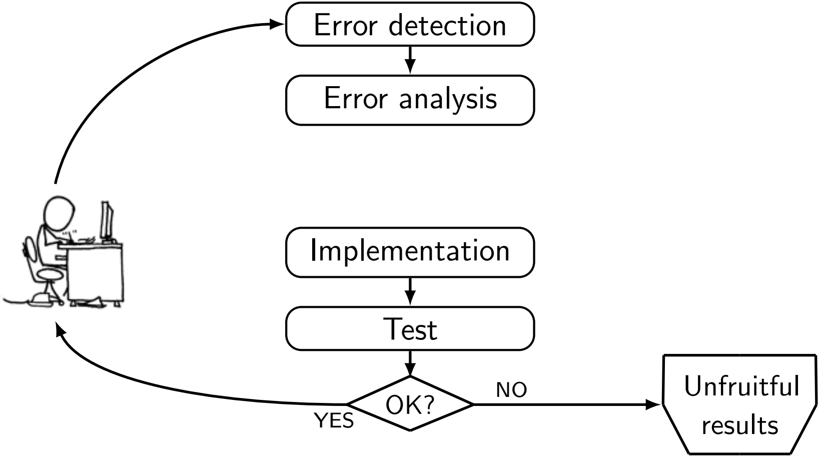
NO

YES

Unfruitful results

Motivation

Evaluation during system development



Error detection

Error analysis

Implementation

Test

OK?

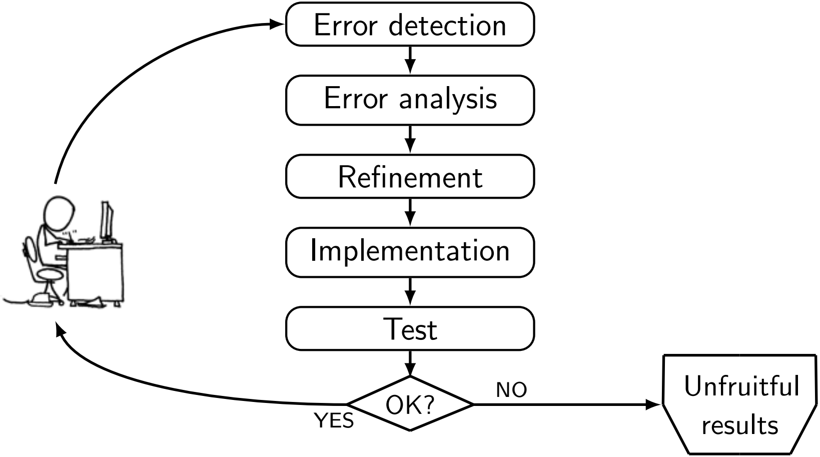
YES

NO

Unfruitful results

Motivation

Evaluation during system development



Error detection

Error analysis

Refinement

Implementation

Test

OK?

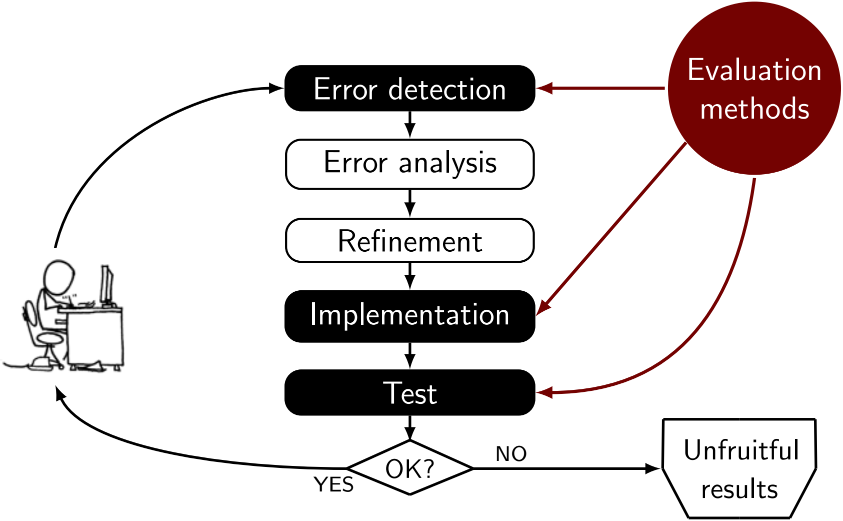
YES

NO

Unfruitful results

Motivation

Evaluation during system development



Motivation

Need for an automatic evaluation

Automatic metrics notably **accelerate** the development cycle of MT systems:

- **Error analysis**
- **System optimisation**
- **System comparison**

Motivation

Need for an automatic evaluation

Automatic metrics notably **accelerate** the development cycle of MT systems:

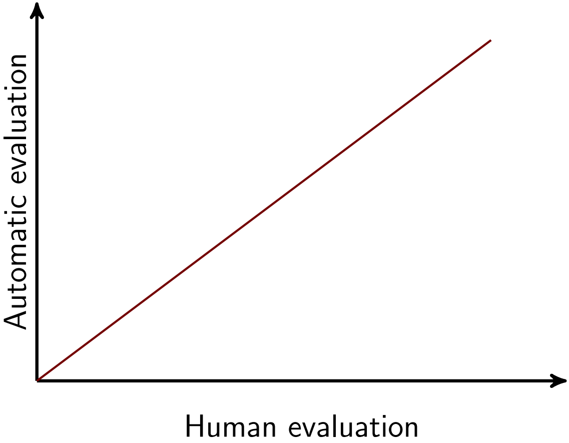
- **Error analysis**
- **System optimisation**
- **System comparison**

Besides, they are

- **Costless** (vs. costly)
- **Objective** (vs. subjective)
- **Reusable** (vs. non-reusable)

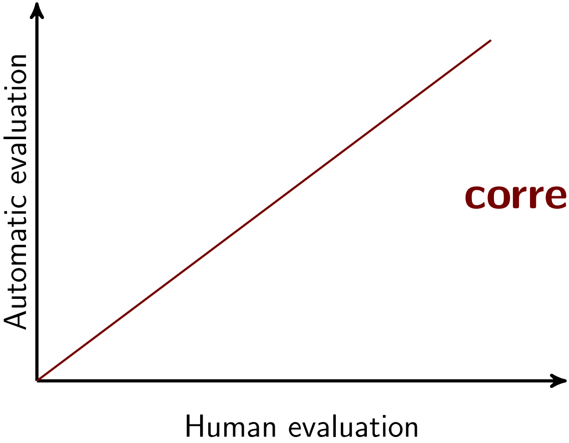
Motivation

But...



Motivation

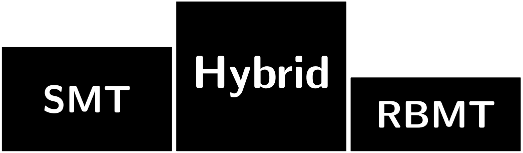
But...



correlate?

Motivation

Our System Evaluation



SMT

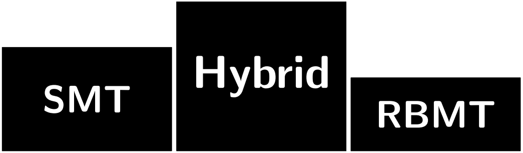
Hybrid

RBMT

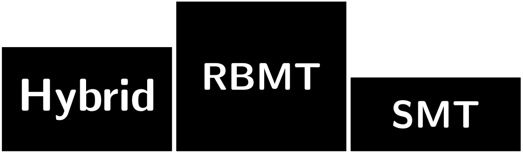
Automatic evaluation

Motivation

Our System Evaluation



Automatic evaluation



Manual evaluation

Motivation

Problem

System development using a metric that does not correlate with human ranking

Are we worsening the system?

SMatxinT, a hybrid translator

Overview

1 Motivation

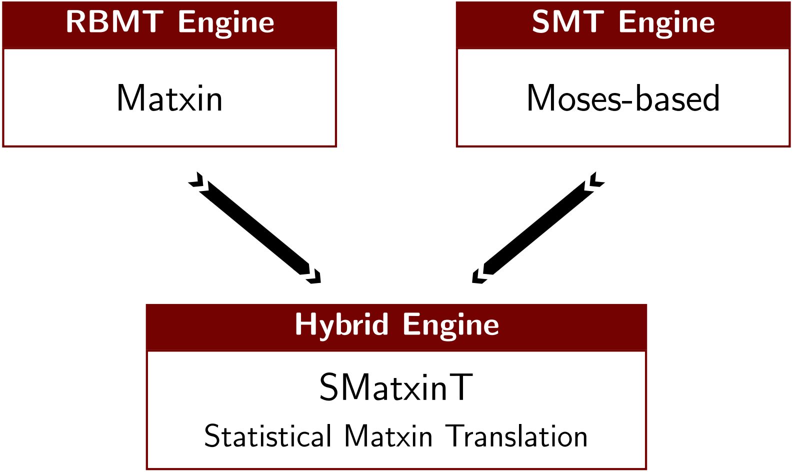
2 SMatxinT, a hybrid translator
- Individual systems
- Hybrid system

3 Systems' evaluation

4 Conclusions

SMatxinT, a hybrid translator

Hybrid engine components



RBMT Engine

Matxin

SMT Engine

Moses-based

Hybrid Engine

SMatxinT

Statistical Matxin Translation

SMatxinT, a hybrid translator

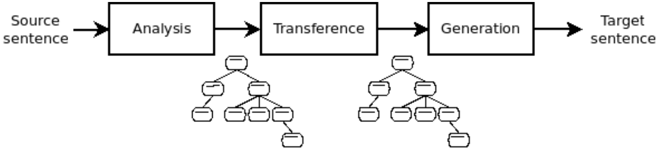
SMT System

- **Basque segmentation**
- **Language model**: 3-gram interpolated Kneser-Ney discounting, SRILM Toolkit
- **Alignments**: GIZA++ Toolkit
- **Translation model**: Moses package
- **Weights optimization**: MERT against BLEU
- **Decoder**: Moses

SMatxinT, a hybrid translator

RBMT System

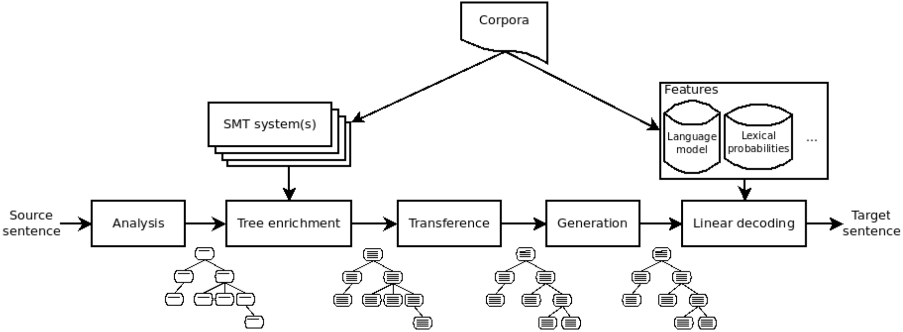
Matxin, Open-Source Rule-Based MT system



- Chunk-based dependency tree
  (Dependency trees + chunk boundaries)

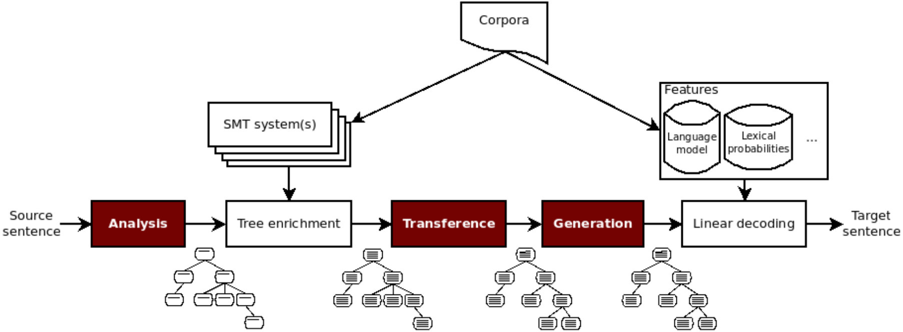
SMatxinT, a hybrid translator

SMatxinT System

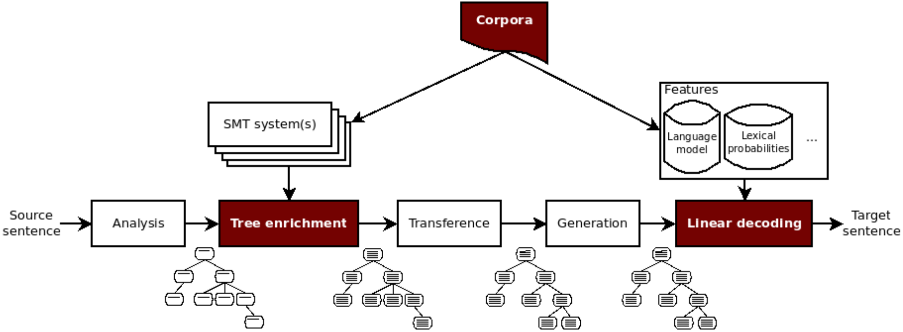


SMatxinT, a hybrid translator

SMatxinT System

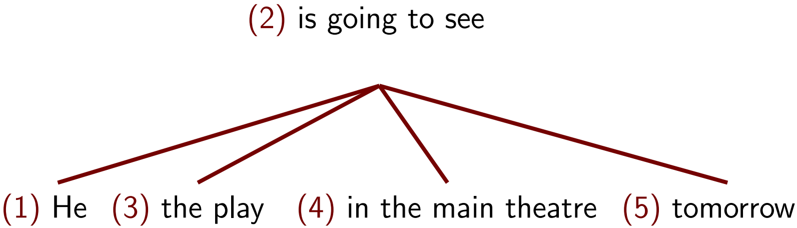


SMatxinT System



SMatxinT, a hybrid translator

Tree enrichment

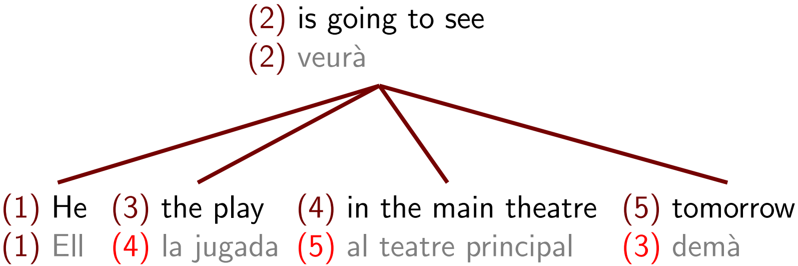


(2) is going to see

(1) He (3) the play (4) in the main theatre (5) tomorrow

SMatxinT, a hybrid translator

Tree enrichment



(2) is going to see
(2) veurà

(1) He
(1) Ell

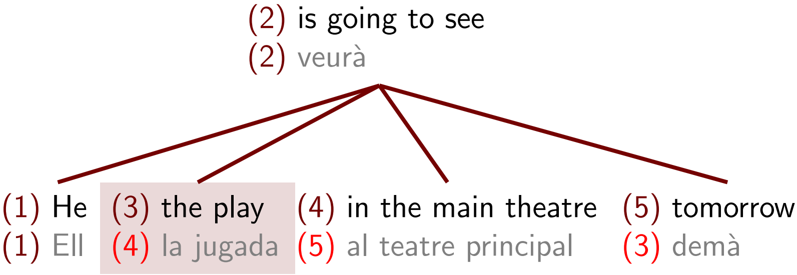
(3) the play
(4) la jugada

(4) in the main theatre
(5) al teatre principal

(5) tomorrow
(3) demà

SMatxinT, a hybrid translator

Tree enrichment



(2) is going to see
(2) veurà

(1) He
(1) Ell

(3) the play
(4) la jugada

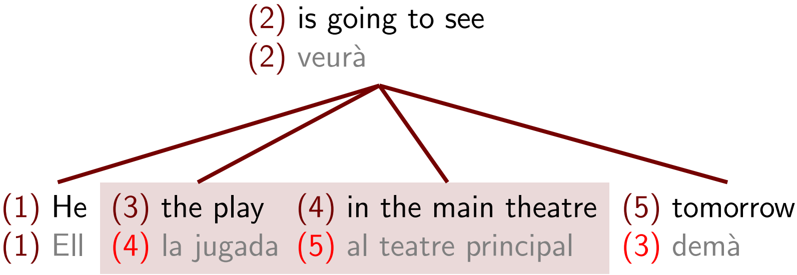
(4) in the main theatre
(5) al teatre principal

(5) tomorrow
(3) demà

SMT: l'obra
...

SMatxinT, a hybrid translator

Tree enrichment



(2) is going to see
(2) veurà

(1) He
(1) Ell

(3) the play
(4) la jugada

(4) in the main theatre
(5) al teatre principal

(5) tomorrow
(3) demà

SMT: l'obra

SMT: l'obra
l'obra al cinema principal
l'obra al teatre principal
...

SMatxinT, a hybrid translator

Tree enrichment: purpose

The RBMT system

- ensures syntactic correctness,
- and takes care of long distance reordering.

Additional richness of phrases:

- Short phrases to improve lexical selection
- Long phrases to overcome wrong syntactic analysis

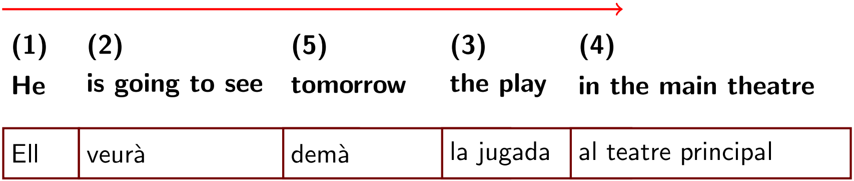
SMatxinT, a hybrid translator

Linear decoding

(1)	**(2)**	**(5)**	**(3)**	**(4)**
He	**is going to see**	**tomorrow**	**the play**	**in the main theatre**

SMatxinT, a hybrid translator

Linear decoding



(1)	(2)	(5)	(3)	(4)
He	**is going to see**	**tomorrow**	**the play**	**in the main theatre**
Ell	veurà	demà	la jugada	al teatre principal

SMatxinT, a hybrid translator

Linear decoding

(1) **He**	**(2)** **is going to see**	**(5)** **tomorrow**	**(3)** **the play**	**(4)** **in the main theatre**
Ell	veurà	demà	la jugada	al teatre principal
Ell ϕ	veurà mirarà ...	demà	l’obra la jugada ...	al teatre principal al cinema principal al teatre del centre

SMatxinT, a hybrid translator

Linear decoding

(1)	**(2)**	**(5)**	**(3)**	**(4)**
He	**is going to see**	**tomorrow**	**the play**	**in the main theatre**
Ell	veurà	demà	la jugada	al teatre principal
Ell ϕ	veurà mirarà ...	demà	l'obra la jugada ...	al teatre principal al cinema principal al teatre del centre
Ell ϕ	veurà mirarà ...	demà	l'obra al cinema del centre l'obra al teatre principal ...	

SMatxinT, a hybrid translator

Linear decoding

(1)	(2)	(5)	(3)	(4)
He	**is going to see**	**tomorrow**	**the play**	**in the main theatre**
Ell	veurà	demà	la jugada	al teatre principal
Ell ϕ	veurà mirarà ...	demà	l'obra la jugada ...	al teatre principal al cinema principal al teatre del centre
Ell ϕ	veurà mirarà ...	demà	l'obra al cinema del centre l'obra al teatre principal ...	

...

Anirà a veure demà l'obra al teatre principal
Ell mirarà demà la jugada al teatre principal
...

SMatxinT, a hybrid translator

Linear decoding: Features

Standard SMT features

- Language model
- Word penalty
- Phrase penalty

SMatxinT, a hybrid translator

Linear decoding: Features

Standard SMT features

- Language model
- Word penalty
- Phrase penalty

Source/consensus features

- Counter ($1 \ldots n$)
- SMT ($1/e$)
- RBMT ($1/e$)
- Both ($e^{\#}$)

SMatxinT, a hybrid translator

Linear decoding: Features

Standard SMT features

- Language model
- Word penalty
- Phrase penalty

Source/consensus features

- Counter (1...n)
- SMT (1/e)
- RBMT (1/e)
- Both ($e^{\#}$)

Lexical features

- Corpus lexical probabilities (eu2es & es2eu)
- Dictionary lexical probabilities (eu2es & es2eu)

SMatxinT, a hybrid translator

The log-linear model & MERT

Features, $h_m(f, e)$

$$\hat{e} = \mathrm{argmax_e} \log P(e|f) = \mathrm{argmax_e} \sum \lambda_m \, h_m(f, e)$$

MERT

λ_m that minimise $\Delta_{err} \equiv \texttt{metric}(\hat{e})\text{-}\texttt{metric}(e_{ref})$

Systems' evaluation

Overview

1. Motivation

2. SMatxinT, a hybrid translator

3. Systems' evaluation
   - Setting
   - Automatic evaluation
   - Manual evaluation

4. Conclusions

Systems' evaluation

Corpora

Language pair

- Spanish–Basque

Training corpus

- Administrative documents and TV programs descriptions
- 491,853 parallel sentences

Development and test corpora

- *Elhuyar dev&test:* Administrative documents (1500 snt)
- *NEWS:* News (1500 sentences, 2 references)

Systems' evaluation

Systems

Individual systems

- SMT
- Matxin

Systems' evaluation

Systems

Individual systems

- SMT
- Matxin

Hybrid systems

- SMatxinT with different MERT

λ_m that minimise $\quad \texttt{metric}(\hat{e})\text{-}\texttt{metric}(e_{ref})$

$\texttt{metric}$: BLEU, BLEU$_C$, METEOR

Systems' evaluation

Systems

Individual systems

- SMT
- Matxin

Hybrid systems

- SMatxinT with different MERT

λ_m that minimise $\quad \texttt{metric}(\hat{e})\text{-}\texttt{metric}(e_{ref})$

$\texttt{metric}$: BLEU, **BLEU$_C$**, METEOR

Systems' evaluation

Systems

Hybrid systems

- SMatxinT with different MERT

λ_m that minimise $\texttt{metric}(\hat{e})\text{-}\texttt{metric}(e_{ref})$

`metric`: BLEU, BLEU$_C$, METEOR

$$\text{BLEU}_C = (\text{BLEU} + \text{BLEU}_{PoS})/2$$

Systems' evaluation

Systems

Hybrid systems

- SMatxinT with different MERT

λ_m that minimise $\texttt{metric}(\hat{e})\text{-}\texttt{metric}(e_{ref})$

$\texttt{metric}$: BLEU, BLEU$_C$, METEOR

$$\text{BLEU}_C = (\text{BLEU} + \text{BLEU}_{PoS})/2$$

Control system

- Google

Systems' evaluation

In-domain automatic evaluation

	BLEU	**METEOR**	**TER**	**BLEU$_c$**
Matxin	6.07	27.20	83.49	19.65
SMT	**16.50**	37.49	70.39	27.64

Systems' evaluation

In-domain automatic evaluation

	BLEU	**METEOR**	**TER**	**BLEU$_c$**
Matxin	6.07	27.20	83.49	19.65
SMT	**16.50**	37.49	70.39	27.64
Google	8.19	28.02	78.43	20.73
SMatxinT$_{\text{BL}}$	16.09	**38.24**	69.92	**27.95**
SMatxinT$_{\text{BL}_c}$	15.36	**38.24**	70.78	27.33
SMatxinT$_{\text{MTR}}$	15.87	37.77	**67.77**	27.53

Systems' evaluation

Out-of-domain automatic evaluation

	BLEU	**METEOR**	**TER**	**BLEU$_c$**
Matxin	12.67	36.10	69.16	**31.98**
SMT	15.84	37.70	66.52	**31.01**

Systems' evaluation

Out-of-domain automatic evaluation

	BLEU	**METEOR**	**TER**	**BLEU$_c$**
Matxin	12.67	36.10	69.16	31.98
SMT	15.84	37.70	66.52	31.01
Google	12.36	32.57	70.44	29.08
SMatxinT$_{\text{BL}}$	16.61	39.24	64.50	32.77
SMatxinT$_{\text{BL}_c}$	**17.11**	**39.94**	63.84	**33.39**
SMatxinT$_{\text{MTR}}$	16.76	39.30	**62.83**	32.50

Systems' evaluation

Manual evaluation

- 100 sentences in-domain, 100 sentences out-of domain
- 2 evaluators for each sentence

1st. experiment: 5 systems to rank per sentence

8.: **El Supremo ordena juzgar a patrones de cayucos interceptados en alta mar**

( ) ⇒ gorenak epaitzeko agindu dio eredu cayucos interceptados en alta mar
( ) ⇒ gorenak cayucos-interceptados eredu itsaso zabalean epaitzeko agindu
( ) ⇒ Supremok cayucos geldituen patroiak itsas zabalean epaitzea ordenatzen du
( ) ⇒ gorenak cayucos-interceptados eredu en alta mar epaitzeko agindu du
( ) ⇒ gorenak epaitzeko agindu dio patrones de cayucos interceptados en alta mar

Systems' evaluation

Manual evaluation

- Ranking allows ties

1.: **La oposición cree que el unico relevo necesario es el del alcalde**

( ) ⇒ oposizioaren ustez beharrezkoa da txanda bakarra alkatearen
( ) ⇒ oposizioak sinesten du beharrezko txanda bakarra alkatetik dela
( ) ⇒ oposizioaren ustez beharrezkoa da txanda bakarra alkatearen
( ) ⇒ oposizioaren ustez beharrezkoa da txanda bakarra alkatearen
( ) ⇒ oposizioaren ustez beharrezkoa da txanda bakarra alkatearen

Systems' evaluation

Manual evaluation

- Ranking allows ties

1.: **La oposición cree que el unico relevo necesario es el del alcalde**

( ) ⇒ oposizioaren ustez beharrezkoa da txanda bakarra alkatearen
( ) ⇒ oposizioak sinesten du beharrezko txanda bakarra alkatetik dela
( ) ⇒ oposizioaren ustez beharrezkoa da txanda bakarra alkatearen
( ) ⇒ oposizioaren ustez beharrezkoa da txanda bakarra alkatearen
( ) ⇒ oposizioaren ustez beharrezkoa da txanda bakarra alkatearen

- Two measures:

  Ranking mean, from [1,5] to [1,1] if ties
  Normalisation to [0,1]

Systems' evaluation

Manual evaluation

	in-domain		out-of-domain	
	Rank	**Norm**	**Rank**	**Norm**
Matxin	**2.07**	**0.396**	**1.70**	**0.275**
SMT	2.51	0.532	2.60	0.625

Systems' evaluation

Manual evaluation

	in-domain		out-of-domain	
	Rank	**Norm**	**Rank**	**Norm**
Matxin	**2.07**	**0.396**	**1.70**	**0.275**
SMT	2.51	0.532	2.60	0.625
SMatxinT$_{\mathbf{BL}}$	2.16	0.423	2.21	0.485
SMatxinT$_{\mathbf{BL}_c}$	2.08	0.399	2.11	0.445
SMatxinT$_{\mathbf{MTR}}$	2.09	0.403	2.12	0.470

Systems' evaluation

Manual evaluation

2nd experiment: Discrete ranking (instead of mean values)

Each system is qualified for each sentence as

best

intermediate

worst

all-draw

Systems' evaluation

Manual evaluation

in-domain

	Best	Intermediate	Worst	All-draw
Matxin	**24** (34+42)	9 (26+19)	20 (38+32)	0 (2+7)
SMT	9 (22+23)	7 (31+23)	**30** (45+47)	0 (2+7)
SMatxinT$_{\mathbf{BL}}$	8 (27+19)	22 (52+43)	8 (19+31)	0 (2+7)
SMatxinT$_{\mathbf{BL}_c}$	12 (27+18)	**29** (55+45)	7 (16+30)	0 (2+7)
SMatxinT$_{\mathbf{MTR}}$	6 (28+19)	24 (54+47)	6 (16+27)	0 (2+7)

Systems' evaluation

Manual evaluation

out-of-domain

	Best	**Intermediate**	**Worst**	**All-draw**
Matxin	**47** (51+64)	4 (22+12)	10 (25+19)	0 (2+5)
SMT	7 (20+11)	6 (21+25)	**41** (57+59)	0 (2+5)
SMatxinT$_{\mathbf{BL}}$	11 (28+15)	27 (44+43)	21 (26+37)	0 (2+5)
SMatxinT$_{\mathbf{BL}_c}$	12 (27+17)	**28** (50+44)	15 (21+34)	0 (2+5)
SMatxinT$_{\mathbf{MTR}}$	11 (26+16)	26 (46+42)	18 (26+37)	0 (2+5)

Conclusions

Comments & summary

System evaluation

- Evaluation and comparison of SMT, Matxin and SMatxinT.
- Human evaluation and lexical metrics do not correlate in this case.
- BLEU$_c$ does a slightly better job than BLEU.

Conclusions

Comments & summary

System development

- Development (MERT) should use metrics that correlate with human assessments in order to improve translations.
- SMatxinT$_{\mathrm{BL}_c}$ *b*est hybrid, but minimal differences.

Conclusions

Comments & summary

System development

- Development (MERT) should use metrics that correlate with human assessments in order to improve translations.
- SMatxinT$_{\mathrm{BL}_c}$ *b*est hybrid, but **minimal differences**.
- **Why?**

Conclusions

Comments & summary

MERT development

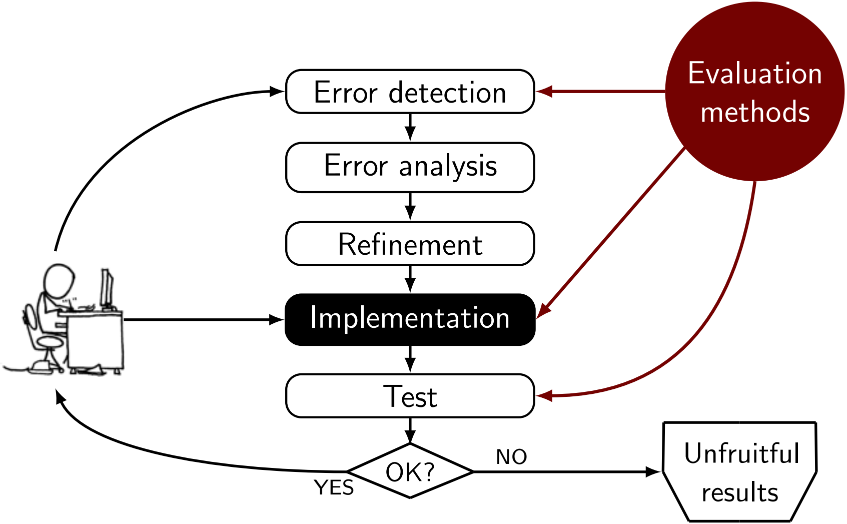
- We do not provide MERT with features sensible to the metric.

Next steps

- Include more linguistic features in SMatxinT.
- Define a metric that includes these features to be used with MERT.

Conclusions

Current and next steps



Error detection

Error analysis

Refinement

Implementation

Test

OK?

YES

NO

Unfruitful results

Evaluation methods

Conclusions

Thanks for your attention

Thank you!

Deep evaluation of hybrid architectures: Use of different metrics in MERT weight optimization

Cristina España-Bonet, Gorka Labaka,
Arantza Díaz de Ilarraza, Kepa Sarasola, Lluís Màrquez

Free/Open-source Rule-based Machine Translation

Gothenburg, June 14th, 2012