Hybrid Machine Translation Guided by a Rule–Based System

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

Universitat Politècnica de Catalunya, **University of the Basque Country**

Motivation

- All the current MT approaches have their pros and cons.
- Rule-Based MT
- Statistical MT
- We would like to get the best of each world.

Motivation

- All the current MT approaches have their pros and cons.
- Rule-Based MT
  - \+ Syntactically better translations: long distance reordering, agreement
  - – Worse lexical selection
  - – Performance degradation for unexpected syntactic structures
- Statistical MT
- We would like to get the best of each world.

Motivation

- All the current MT approaches have their pros and cons.
- Rule-Based MT : +syntax, −lexical selection, −unexpected structures
- Statistical MT
  - \+ Better lexical selection and fluency
  - – Structurally worse translations
  - – Performance degradation for *out-domain* texts.
- We would like to get the best of each world.

Motivation

- All the current MT approaches have their pros and cons.
- Rule-Based MT : +syntax, −lexical selection, −unexpected structures
- Statistical MT : +lexical selection, −long distance reordering, −out-domain performance
- We would like to get the best of each world.
  - RBMT's grammatical correctness.
  - SMT's lexical selection
  - SMT tolerance to unexpected structures.

Outline

1. SMatxinT: RBMT leaded hybrid system
2. Experimental Results
3. Conclusions

Outline

Statistical machine translator(s)

- SMTb (Moses with default configuration):
  - phrase and lexical translation probabilities
  - 3-gram LM
  - lexicalized reordering
  - ...

Statistical machine translator(s)

- SMTb (Moses with default configuration):
  - phrase and lexical translation probabilities
  - 3-gram LM
  - lexicalized reordering
  - ...
- SMTm:
  - Uses segmentation of Basque
  - Default Moses configuration: phrase and lexical prob., 3-gram LM, ...
  - Word generation phase to go from segmented text to final words
  - Word level LM added using n-best list

Statistical machine translator(s)

- SMTb (Moses with default configuration):
  - phrase and lexical translation probabilities
  - 3-gram LM
  - lexicalized reordering
  - ...
- SMTm:
  - Uses segmentation of Basque
  - Default Moses configuration: phrase and lexical prob., 3-gram LM, ...
  - Word generation phase to go from segmented text to final words
  - Word level LM added using n-best list
- ...

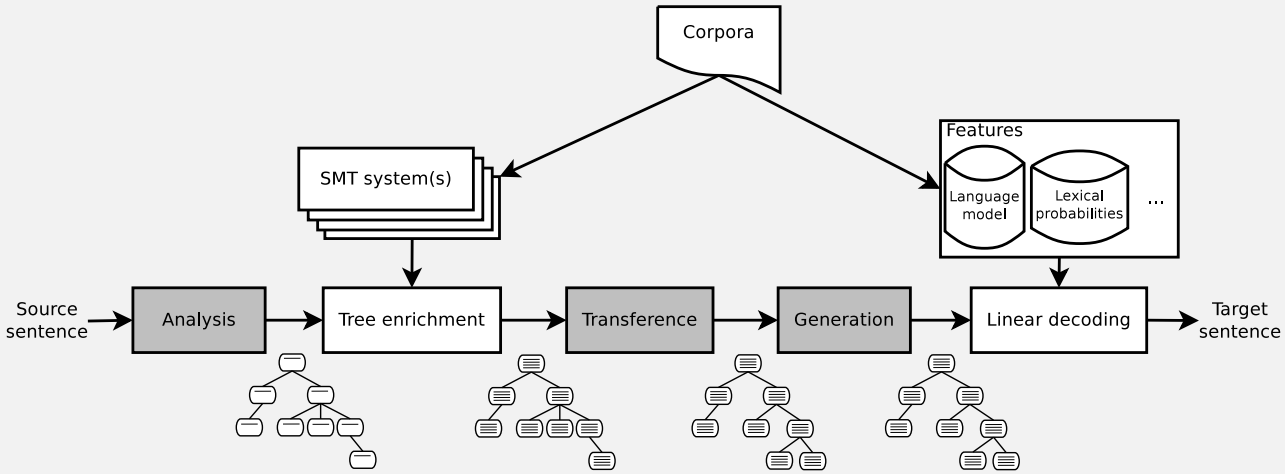
Matxin, a rule-based translator

- In-house developed Open-Source Rule-Based MT system.
- Classical transfer-based approach: analysis, transfer and generation.
- Chunk-based dependency tree:
  - Dependency trees + chunk boundaries.

SMatxinT: RBMT guided Hybrid MT

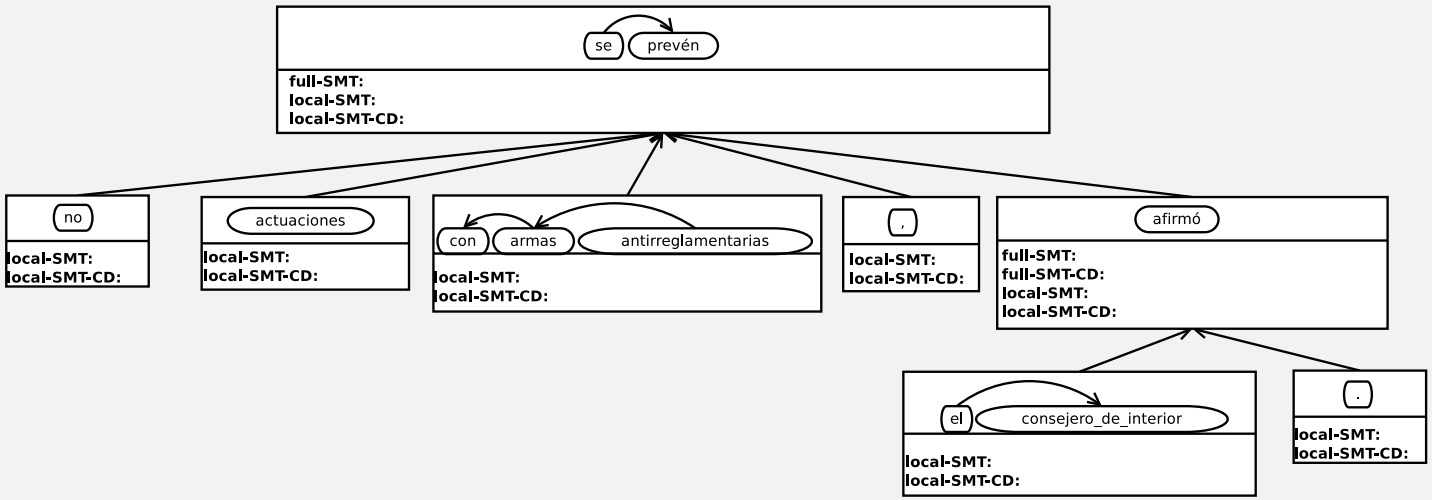
- Translation process is guide by the Rule-Based system.
  - Ensures syntactic correctness.
  - Takes care of long distance reordering.
- Allow substitution of RBMT partial translations with their SMT counterparts.
  - Substitutions of short strings improve lexical selection.
  - Longer substitutions allow to overcome wrong syntactic analysis.

SMatxinT architecture



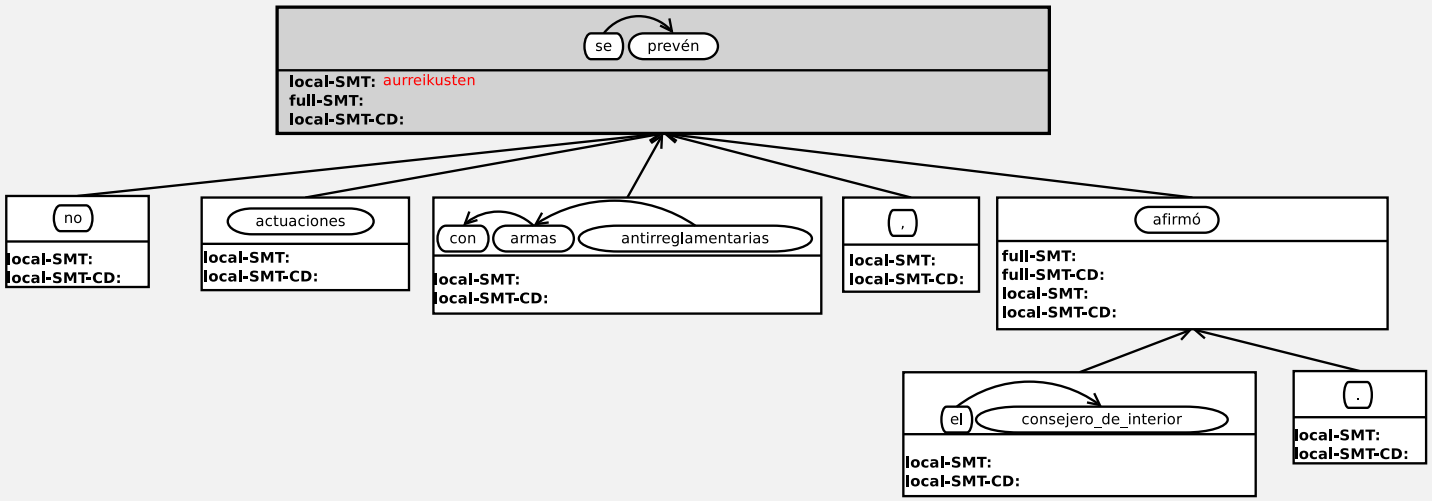
- Two new modules are added to the RBMT architecture.
  - Tree enrichment (between analysis and transfer).
  - linear decoding (after generation).

Tree enrichment



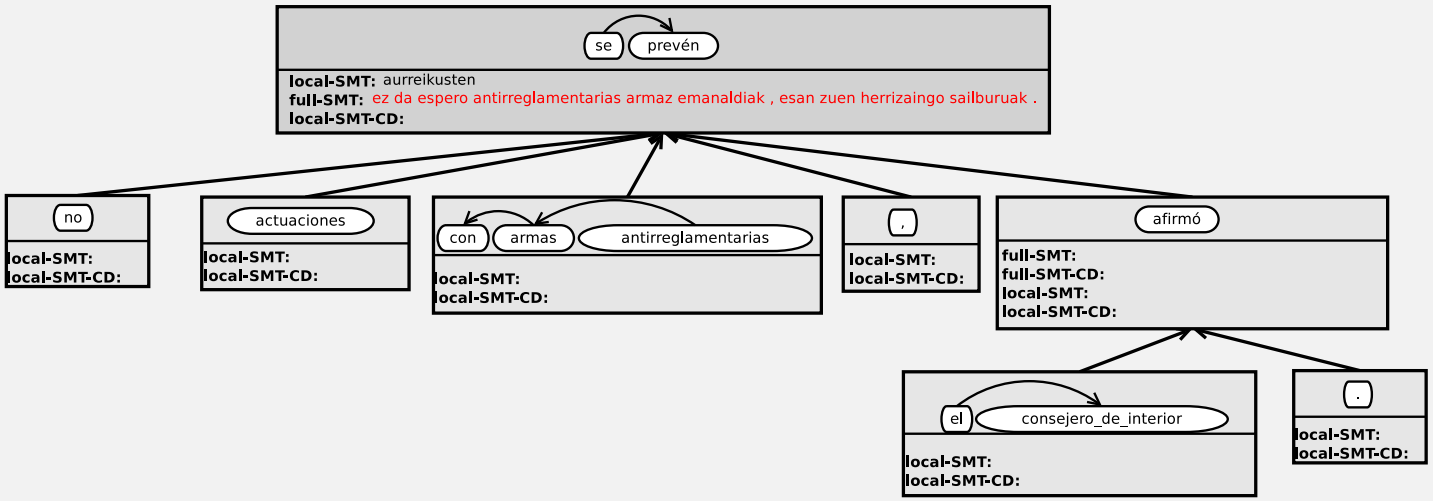
- After analysis, and before transfer.
- Each phrase in the tree is enrich with one (or several) SMT translation counterpart.
- Two types of SMT correspondences:

Tree enrichment



- After analysis, and before transfer.
- Each phrase in the tree is enrich with one (or several) SMT translation counterpart.
- Two types of SMT correspondences:
  - local: Allows SMatxinT to use SMT lexical selection

Tree enrichment



- After analysis, and before transfer.
- Each phrase in the tree is enrich with one (or several) SMT translation counterpart.
- Two types of SMT correspondences:
  - local: Allows SMatxinT to use SMT lexical selection
  - full subtree: Allows SMatxinT to overcome analysis errors.

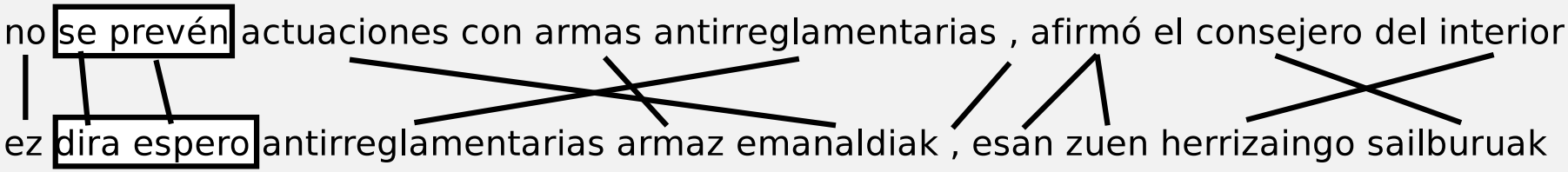
Tree enrichment: Context Discriminate translations

- Many chunks are too small to SMT
  - Few context to get proper translation.
- We also extract corresponding local translation form longer SMT translations.
  - Based on alignment provided by SMT.
  - Allows the SMT to use the context.

Tree enrichment: Context Discriminate translations

- Many chunks are too small to SMT
  - Few context to get proper translation.
- We also extract corresponding local translation form longer SMT translations.
  - Based on alignment provided by SMT.
  - Allows the SMT to use the context.

Tree enrichment: Context Discriminate translations



- Many chunks are too small to SMT
  - Few context to get proper translation.
- We also extract corresponding local translation form longer SMT translations.
  - Based on alignment provided by SMT.
  - Allows the SMT to use the context.

Linear decoding

- After tree enrichment, transfer and generation are applied as usual.
- RBMT translation enriched with several candidates for each phrase.

emanaldiak ez dituzte aurreikusten arauz kontrako armekin , barne sailburua baieztetu zuen

Linear decoding

- After tree enrichment, transfer and generation are applied as usual.
- RBMT translation enriched with several candidates for each phrase.

emanaldiak	ez	dituzte aurreikusten	arauz kontrako armekin	,	barne sailburua	baieztetu zuen
jarduera	ez	aurreikusten	antirreglamentarias armaz	,	barne sailburua	esan zuen
emanaldiak	ez	da espero	antirreglementarias armaz	,	herrizaingo zailburuak	esan zuen

Linear decoding

- After tree enrichment, transfer and generation are applied as usual.
- RBMT translation enriched with several candidates for each phrase.

emanaldiak	ez	dituzte aurreikusten	arauz kontrako armekin	,	barne sailburua	baieztetu zuen
jarduera	ez	aurreikusten	antirreglamentarias armaz	,	barne sailburua	esan zuen
emanaldiak	ez	da espero	antirreglementarias armaz	,	herrizaingo zailburuak	esan zuen

esan zuen barne sailburuak

ez da espero antirreglamentarias armaz emanldiak , esan zuen herrizaingo sailburuak

Linear decoding

- After tree enrichment, transfer and generation are applied as usual.
- RBMT translation enriched with several candidates for each phrase.
- A Linear decoding module is used to choose the best candidate for each phrase.

emanaldiak	ez	dituzte aurreikusten	arauz kontrako armekin	,	barne sailburua	baieztetu zuen
jarduera	ez	aurreikusten	antirreglamentarias armaz	,	barne sailburua	esan zuen
emanaldiak	ez	da espero	antirreglementarias armaz	,	herrizaingo zailburuak	esan zuen

esan zuen barne sailburuak

ez da espero antirreglamentarias armaz emanldiak , esan zuen herrizaingo sailburuak

Linear decoding

- After tree enrichment, transfer and generation are applied as usual.
- RBMT translation enriched with several candidates for each phrase.
- A Linear decoding module is used to choose the best candidate for each phrase.
  - We used Moses (in monotonous way) as linear decoder.
  - A wide range or features are defined.

emanaldiak	ez	dituzte aurreikusten	arauz kontrako armekin	,	barne sailburua	baieztetu zuen
jarduera	ez	aurreikusten	antirreglamentarias armaz	,	barne sailburua	esan zuen
emanaldiak	ez	da espero	antirreglementarias armaz	,	herrizaingo zailburuak	esan zuen

esan zuen barne sailburuak

ez da espero antirreglamentarias armaz emanldiak , esan zuen herrizaingo sailburuak

Linear decoding: Features

Standard SMT features

- Language model
- Word penalty
- Phrase penalty

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 ($e^{\#}$)

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)

Outline

1. SMatxinT: RBMT leaded hybrid system

2. Experimental Results
   - Corpora
   - Systems
   - Results

3. Conclusions

Corpora

Language pair

- Spanish–Basque

Training corpus

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

Development and test corpora

- *Elhuyar dev & test:* Administrative documents (1500 sentences)
- *EITB:* News (1500 sentences, 1 reference)
- *NEWS:* News (1500 sentences, 2 references)

Systems

Individual systems

- SMTb
- SMTm
- Matxin

Systems

Individual systems

- SMTb
- SMTm
- Matxin

Hybrid systems

- SMatxinT$_0$: Hybrid system where only SMT translations are used
- SMatxinT

Systems

Individual systems

- SMTb
- SMTm
- Matxin

Hybrid systems

- SMatxinT$_0$
- SMatxinT

Control system

- Google

Automatic Metrics

	Elhuyar		EITB		NEWS	
	BLEU	**TER**	**BLEU**	**TER**	**BLEU**	**TER**
Matxin	5.10	83.32	5.77	87.73	11.72	82.04
SMTb	14.96	70.20	8.03	83.27	14.74	78.63
SMTm	13.71	71.64	7.64	85.59	14.58	78.90
Google	7.32	78.43	6.73	86.32	12.01	81.84
SMatxinT$_0$	14.50	69.73	8.45	82.17	14.90	77.29
SMatxinT	14.73	69.18	8.81	81.33	15.31	76.54

Automatic Metrics

	Elhuyar		EITB		NEWS	
	BLEU	**TER**	**BLEU**	**TER**	**BLEU**	**TER**
Matxin	5.10	83.32	5.77	87.73	11.72	82.04
SMTb	14.96	70.20	8.03	83.27	14.74	78.63
SMTm	13.71	71.64	7.64	85.59	14.58	78.90
Google	7.32	78.43	6.73	86.32	12.01	81.84
SMatxinT$_0$	14.50	69.73	8.45	82.17	14.90	77.29
SMatxinT	14.73	69.18	8.81	81.33	15.31	76.54

BLEU and TER scores are generally low

Automatic Metrics

	Elhuyar		EITB		NEWS	
	BLEU	**TER**	**BLEU**	**TER**	**BLEU**	**TER**
Matxin	5.10	83.32	5.77	87.73	11.72	82.04
SMTb	14.96	70.20	8.03	83.27	14.74	78.63
SMTm	13.71	71.64	7.64	85.59	14.58	78.90
Google	7.32	78.43	6.73	86.32	12.01	81.84
SMatxinT$_0$	14.50	69.73	8.45	82.17	14.90	77.29
SMatxinT	14.73	69.18	8.81	81.33	15.31	76.54

All systems outperform the RBMT

Automatic Metrics

	Elhuyar		EITB		NEWS	
	BLEU	**TER**	**BLEU**	**TER**	**BLEU**	**TER**
Matxin	5.10	83.32	5.77	87.73	11.72	82.04
SMTb	14.96	70.20	8.03	83.27	14.74	78.63
SMTm	13.71	71.64	7.64	85.59	14.58	78.90
Google	7.32	78.43	6.73	86.32	12.01	81.84
SMatxinT$_0$	14.50	69.73	8.45	82.17	14.90	77.29
SMatxinT	14.73	69.18	8.81	81.33	15.31	76.54

Results of hybrid system are comparable but not better than SMT subsystems

Automatic Metrics

	Elhuyar		EITB		NEWS	
	BLEU	TER	BLEU	TER	BLEU	TER
Matxin	5.10	83.32	5.77	87.73	11.72	82.04
SMTb	14.96	70.20	8.03	83.27	14.74	78.63
SMTm	13.71	71.64	7.64	85.59	14.58	78.90
Google	7.32	78.43	6.73	86.32	12.01	81.84
SMatxinT$_0$	14.50	69.73	8.45	82.17	14.90	77.29
SMatxinT	14.73	69.18	8.81	81.33	15.31	76.54

SMatxinT and SMatxinT$_0$ are consistently better than single systems

Automatic Metrics

	Elhuyar		EITB		NEWS	
	BLEU	TER	BLEU	TER	BLEU	TER
Matxin	5.10	83.32	5.77	87.73	11.72	82.04
SMTb	14.96	70.20	8.03	83.27	14.74	78.63
SMTm	13.71	71.64	7.64	85.59	14.58	78.90
Google	7.32	78.43	6.73	86.32	12.01	81.84
SMatxinT$_0$	14.50	69.73	8.45	82.17	14.90	77.29
SMatxinT	14.73	69.18	8.81	81.33	15.31	76.54

SMatxinT results are slightly better than SMatxinT$_0$

Source systems

System	SMatxinT		sBLEU Oracle	
	chunks	tokens	chunks	tokens
SMT	2,682 (44.2%)	11,391 (65.4%)	3,202 (38.4%)	9,043 (51.2%)
SMT-CD	523 (8.6%)	1,737 (10.0%)	779 (9.3%)	1,890 (10.7%)
RBMT	401 (6.6%)	1,279 (7.3%)	969 (11.6%)	2,554 (14.4%)
BOTH	2,454 (40.5%)	3,013 (17.3%)	3,389 (40.6%)	4,192 (23.7%)
Total	6,060 (100%)	17,420 (100%)	8,339 (100%)	17,679 (100%)

Source systems

System	SMatxinT		sBLEU Oracle	
	chunks	tokens	chunks	tokens
SMT	2,682 (44.2%)	11,391 (65.4%)	3,202 (38.4%)	9,043 (51.2%)
SMT-CD	523 (8.6%)	1,737 (10.0%)	779 (9.3%)	1,890 (10.7%)
RBMT	401 (6.6%)	1,279 (7.3%)	969 (11.6%)	2,554 (14.4%)
BOTH	2,454 (40.5%)	3,013 (17.3%)	3,389 (40.6%)	4,192 (23.7%)
Total	6,060 (100%)	17,420 (100%)	8,339 (100%)	17,679 (100%)

Hybrid system use very few RBMT translations

Source systems

System	SMatxinT		sBLEU Oracle	
	chunks	tokens	chunks	tokens
SMT	2,682 (44.2%)	11,391 (65.4%)	3,202 (38.4%)	9,043 (51.2%)
SMT-CD	523 (8.6%)	1,737 (10.0%)	779 (9.3%)	1,890 (10.7%)
RBMT	401 (6.6%)	1,279 (7.3%)	969 (11.6%)	2,554 (14.4%)
BOTH	2,454 (40.5%)	3,013 (17.3%)	3,389 (40.6%)	4,192 (23.7%)
Total	6,060 (100%)	17,420 (100%)	8,339 (100%)	17,679 (100%)

Oracle uses more RBMT fragments

Source systems

System	SMatxinT		sBLEU Oracle	
	chunks	tokens	chunks	tokens
SMT	2,682 (44.2%)	11,391 (65.4%)	3,202 (38.4%)	9,043 (51.2%)
SMT-CD	523 (8.6%)	1,737 (10.0%)	779 (9.3%)	1,890 (10.7%)
RBMT	401 (6.6%)	1,279 (7.3%)	969 (11.6%)	2,554 (14.4%)
BOTH	2,454 (40.5%)	3,013 (17.3%)	3,389 (40.6%)	4,192 (23.7%)
Total	6,060 (100%)	17,420 (100%)	8,339 (100%)	17,679 (100%)

Oracle uses more RBMT fragments
The fragments are in average shorter

Source systems

System	SMatxinT		sBLEU Oracle	
	chunks	tokens	chunks	tokens
SMT	2,682 (44.2%)	11,391 (65.4%)	3,202 (38.4%)	9,043 (51.2%)
SMT-CD	523 (8.6%)	1,737 (10.0%)	779 (9.3%)	1,890 (10.7%)
RBMT	401 (6.6%)	1,279 (7.3%)	969 (11.6%)	2,554 (14.4%)
BOTH	2,454 (40.5%)	3,013 (17.3%)	3,389 (40.6%)	4,192 (23.7%)
Total	6,060 (100%)	17,420 (100%)	8,339 (100%)	17,679 (100%)

Linear decoding fails in use RBMT correct information.

Human Evaluation

- 100 random sentences from NEWStest
- Preference between SMT and SMatxinT (2 assessment per sentence)

Assessments	SMT	Tied	SMatxinT
All	45 (22.5%)	64 (32.0%)	91 (45.5%)
Agreement	13 (21.3%)	17 (27.9%)	31 (50.8%)

Human Evaluation

- 100 random sentences from NEWStest
- Preference between SMT and SMatxinT (2 assessment per sentence)

Assessments	SMT	Tied	SMatxinT
All	45 (22.5%)	64 (32.0%)	91 (45.5%)
Agreement	13 (21.3%)	17 (27.9%)	31 (50.8%)

Results confirms that SMatxinT overcomes best SMT system.

Outline

1. SMatxinT: RBMT leaded hybrid system
2. Experimental Results
3. Conclusions

Summary

- We present a hybrid machine translation system that combines RBMT with phrase-based SMT.
- The RBMT system leads the translation process and generates the syntactic structure in the target language.
- The SMT system generates multiple candidate translations of any fragment in this tree. A posterior linear decoder selects the best combination to create the final output.
  - Short translation alternatives correct RBMT lexical selection errors.
  - Longer ones allow to overcome syntactic analysis error.

Conclusions

- SMatxinT achieves statistically significant improvements on out-of-domain test sets for the Spanish-to-Basque translation, according automatic metrics.
- This advantage has been corroborated by a manual evaluation conducted on a set of 100 samples
- The analysis of the oracles shows that there is still a large room for improvement
- Oracle translations tend to be composed by more and shorter chunks, and a larger proportion of chunks coming from RBMT.

Future work

- Define new linguistically based features for the linear decoder to identify correct RBMT translations.
- Alleviate the strong dependence of SMatxinT on the initial syntactic parsing incorporating multiple syntactic trees from the side of the rule based system.
- A more detailed manual comparison of the outputs of the different systems
- Broadening the study to other language pairs.

Thanks

Thanks for your attention.

Hybrid Machine Translation Guided by a Rule–Based System

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

Universitat Politècnica de Catalunya, **University of the Basque Country**