Hybrid Machine Translation in the Abu-MaTran project

Víctor M. Sánchez-Cartagena
Prompsit Language Engineering
vmsanchez@prompsit.com

Workshop on Hybridisation of Machine Translation for Irish
April 2016, Dublin, Ireland

Outline

Outline

Introduction

- Marie Curie Industry-Academia Partnerships and Pathways: Increase industrial adoption of MT, provide MT for Croatian by means of:
  - Automatic acquisition of corpora and linguistic resources
  - Pivot techniques
  - Linguistically augmented SMT
  - Diagnostic evaluation

Introduction

- Marie Curie Industry-Academia Partnerships and Pathways: Increase industrial adoption of MT, provide MT for Croatian by means of:
  - Automatic acquisition of corpora and linguistic resources
  - Pivot techniques
  - **Linguistically augmented SMT**
  - Diagnostic evaluation

Objective

Address the data sparseness problem in morphologically rich languages with the help of shallow-transfer rule-based MT

Phrase-based statistical MT

- Translation: TL sentence with highest probability according to a combination of statistical models

Phrase-based statistical MT

- Translation: TL sentence with highest probability according to a combination of statistical models

Example: *the small houses*

Phrase table:

the	*el*	0.5
the	*las*	0.2
small houses	*casas pequeñas*	0.7
small	*medianas*	0.1
specialised	*especializados*	1

Translation hypotheses:

Phrase-based statistical MT

- Translation: TL sentence with highest probability according to a combination of statistical models

Example: *the small houses*

Phrase table:

the	*el*	0.5
the	*las*	0.2
small houses	*casas pequeñas*	0.7
small	*medianas*	0.1
specialised	*especializados*	1

Translation hypotheses:

el casas pequeñas | 0.35

Phrase-based statistical MT

- Translation: TL sentence with highest probability according to a combination of statistical models

Example: *the small houses*

Phrase table:

the	*el*	0.5
the	*las*	0.2
small houses	*casas pequeñas*	0.7
small	*medianas*	0.1
specialised	*especializados*	1

Translation hypotheses:

el casas pequeñas	0.35
el hogar	0.015
las casas pequeñas	0.14
el medianas hogares	0.015

Phrase-based statistical MT

- Translation is the TL sentence with highest probability given the SL sentence according to a combination of statistical models

Example: *the small houses*

Translation hypotheses:

el casas pequeñas	0.35	**0.3**
el hogar	0.015	**0.7**
las casas pequeñas	0.14	**0.6**
el medianas hogares	0.015	**0.2**

Target language model

How likely is that the translation hypothesis occurs in the target language

Phrase-based statistical MT

- Translation is the TL sentence with highest probability given the SL sentence according to a combination of statistical models

Example: *the small houses*

Final score

Combine translation model score, target language model score, and others

Translation hypotheses:

el casas pequeñas	0.35	0.3	**0.3**
el hogar	0.015	0.7	**0.3**
las casas pequeñas	0.14	0.6	**0.45**
el medianas hogares	0.015	0.2	**0.2**

Phrase-based statistical MT

Data sparseness

- SMT systems usually work with surface forms
- Morphologically rich language $\rightarrow$ difficult to observe in the training corpora all the necessary sequences of inflected forms

Example: Europarl English$\rightarrow$Spanish

Source: *The only specialised category which nobody won ...*
SMT: *La única categoría especializados que nadie ganó ...*

Shallow-transfer rule-based MT

- Borrow linguistic information from Apertium **shallow-transfer** RBMT platform
  - Source-language (SL) and target-language (TL) **intermediate representations**: sequence of lemma, part of speech (PoS) and inflection information

Shallow-transfer rule-based MT

- Borrow linguistic information from Apertium **shallow-transfer** RBMT platform
  - Source-language (SL) and target-language (TL) **intermediate representations**: sequence of lemma, part of speech (PoS) and inflection information

the
small
houses

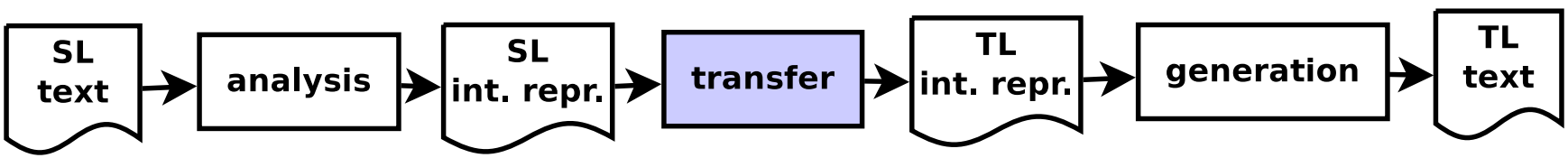
Shallow-transfer rule-based MT

- Borrow linguistic information from Apertium **shallow-transfer** RBMT platform
  - Source-language (SL) and target-language (TL) **intermediate representations**: sequence of lemma, part of speech (PoS) and inflection information

the → *the* `DT`
small → *small* `ADJ`
houses → *house* `N-PL`

Shallow-transfer rule-based MT

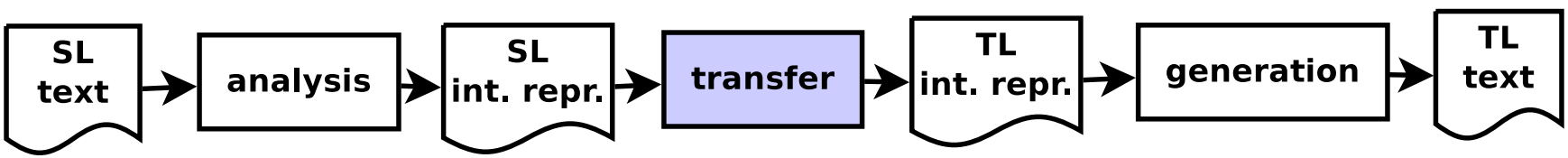
- Borrow linguistic information from Apertium **shallow-transfer** RBMT platform
  - Source-language (SL) and target-language (TL) **intermediate representations**: sequence of lemma, part of speech (PoS) and inflection information



the DT → *el* DT
small ADJ → *pequeño* ADJ
house N-PL → *casa* N-F-PL

Shallow-transfer rule-based MT

- Borrow linguistic information from Apertium **shallow-transfer** RBMT platform
  - Source-language (SL) and target-language (TL) **intermediate representations**: sequence of lemma, part of speech (PoS) and inflection information



el DT		rule:		*el* DT-F-PL
pequeño ADJ	→	ADJ,N to	→	*casa* N-F-PL
casa N-F-PL		N,ADJ + agree		*pequeño* ADJ-F-PL

Shallow-transfer rule-based MT

- Borrow linguistic information from Apertium **shallow-transfer** RBMT platform
  - Source-language (SL) and target-language (TL) **intermediate representations**: sequence of lemma, part of speech (PoS) and inflection information

el `DT-F-PL` → *las*
casa `N-F-PL` → *casas*
pequeño `ADJ-F-PL` → *pequeñas*

Using linguistic information to deal with data sparseness

- Factored translation models

La	única	categoría	especializados	que	nadie	ganó	
DT-F-SG	ADJ-F-SG	N-F-SG	ADJ-M-PL	WDT	PRP	VBD	↓

Using linguistic information to deal with data sparseness

- Factored translation models

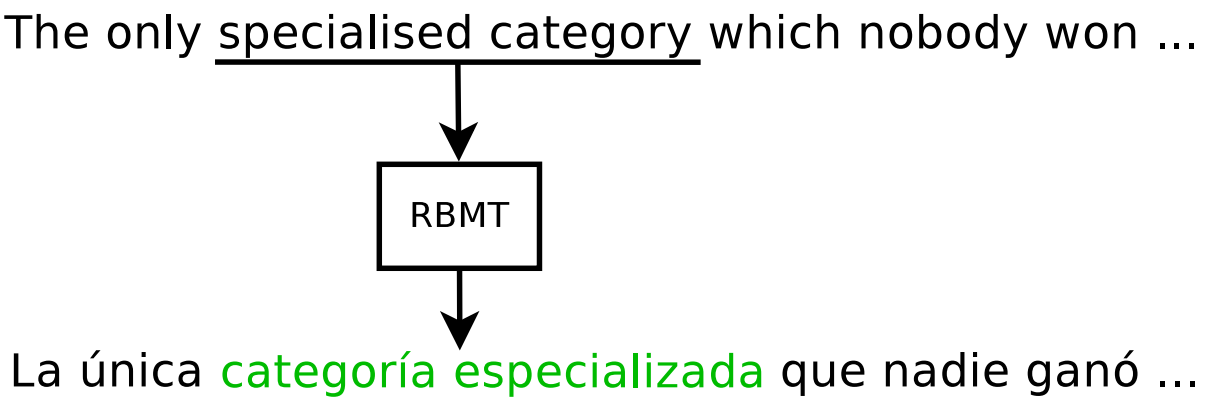
La	única	categoría	especializados	que	nadie	ganó	
DT-F-SG	ADJ-F-SG	N-F-SG	ADJ-M-PL	WDT	PRP	VBD	↓
La	única	categoría	especializada	que	nadie	ganó	
DT-F-SG	ADJ-F-SG	N-F-SG	ADJ-F-SG	WDT	PRP	VBD	↑

Using linguistic information to deal with data sparseness

- Factored translation models

La	única	categoría	especializados	que	nadie	ganó	
DT-F-SG	ADJ-F-SG	N-F-SG	ADJ-M-PL	WDT	PRP	VBD	↓
La	única	categoría	especializada	que	nadie	ganó	
DT-F-SG	ADJ-F-SG	N-F-SG	ADJ-F-SG	WDT	PRP	VBD	↑

- Integrating shallow-transfer rules into SMT

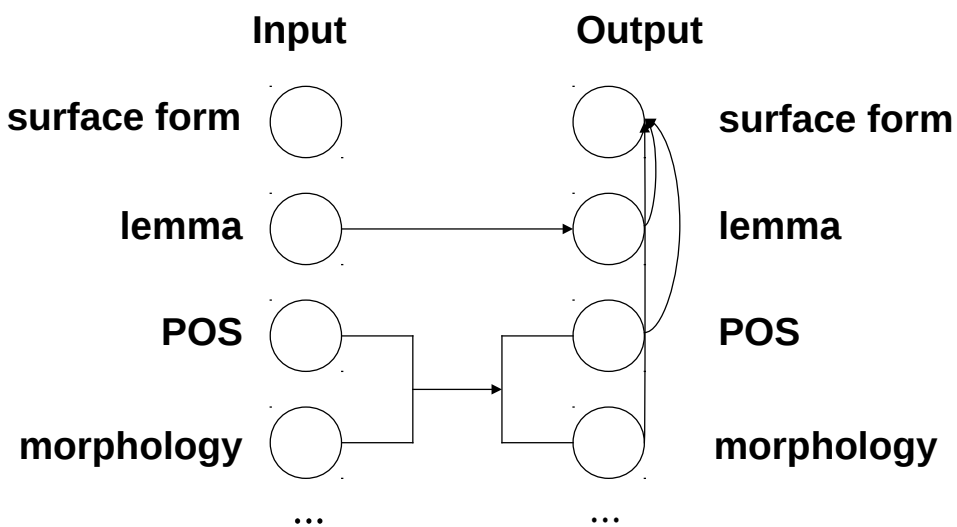


Outline

Definition

Factored translation models:

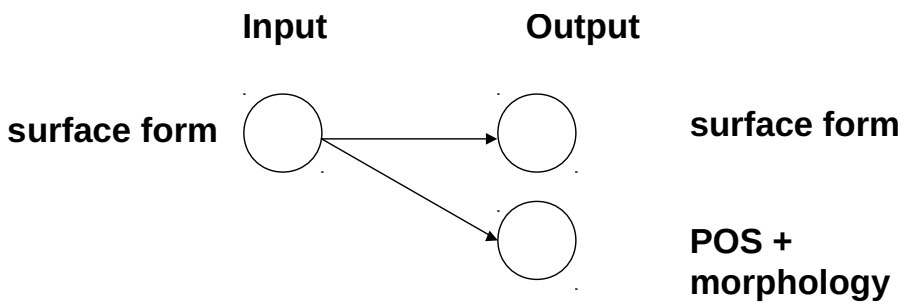
- Extension to surface-form-based SMT:
  - Each word is represented as a set of factors that can be translated independently
  - Final translation may need to be generated from a set of factors by means of an additional (word-level) model.



- Requirement: analysed training corpora

Definition

- In practice, effective and efficient set-up when the TL is highly inflected but the SL is not (Skadins et al., 2010):



Phrase table:

the	*el*|DT-M-SG	0.5
the	*las*|DT-F-PL	0.2
small houses	*casas*|N-F-PL *pequeñas*|ADJ-F-PL	0.7
...	...	...

- Two independent LMs: for surface forms and PoS+morphology tags

Factored models in English to Croatian SMT

- Abu-MaTran **general-domain English-to-Croatian SMT system** from corpora crawled from the Web (1M parallel sentences, 49M Croatian sentences)
- Croatian is a morphologically rich language:
  - Adjectives and nouns inflect for 3 genders, 2 numbers and 7 cases
  - Apertium Croatian morphological lexicon contains more than 900 distinct PoS+morphology tags
- Applied factored models paying attention to:
  - Order of PoS+morphology LM
  - Effect of constraining PoS tagging to lexicon

PoS+morphology language model order

PoS+morphology language model:

La	única	categoría	especializada	que	nadie	ganó
DT-F-SG	ADJ-F-SG	N-F-SG	ADJ-F-SG	WDT	PRP	VBD

$P(\text{sent}) = \prod$

PoS+morphology language model order

PoS+morphology language model:

La	única	categoría	especializada	que	nadie	ganó
DT-F-SG	ADJ-F-SG	N-F-SG	ADJ-F-SG	WDT	PRP	VBD

$$P(\text{sent}) = \prod P(\texttt{DT-F-SG})$$

PoS+morphology language model order

PoS+morphology language model:

La	única	categoría	especializada	que	nadie	ganó
DT-F-SG	ADJ-F-SG	N-F-SG	ADJ-F-SG	WDT	PRP	VBD

$$P(\text{sent}) = \prod P(\texttt{ADJ-F-SG}|\texttt{DT-F-SG})$$

PoS+morphology language model order

PoS+morphology language model:

La	única	categoría	especializada	que	nadie	ganó
DT-F-SG	ADJ-F-SG	N-F-SG	ADJ-F-SG	WDT	PRP	VBD

$$P(\text{sent}) = \prod \quad P(\texttt{N-F-SG}|\texttt{DT-F-SG},\texttt{ADJ-F-SG})$$

PoS+morphology language model order

PoS+morphology language model:

La	única	categoría	especializada	que	nadie	ganó
DT-F-SG	ADJ-F-SG	N-F-SG	ADJ-F-SG	WDT	PRP	VBD

$$P(\text{sent}) = \prod \quad P(\texttt{ADJ-F-SG}|\texttt{DT-F-SG},\texttt{ADJ-F-SG},\texttt{N-F-SG})$$

PoS+morphology language model order

PoS+morphology language model:

La	única	categoría	especializada	que	nadie	ganó
DT-F-SG	ADJ-F-SG	N-F-SG	ADJ-F-SG	WDT	PRP	VBD

$$P(\text{sent}) = \prod \quad P(\texttt{WDT}|\texttt{DT-F-SG},\texttt{ADJ-F-SG},\texttt{N-F-SG},\texttt{ADJ-F-SG})$$

PoS+morphology language model order

PoS+morphology language model:

La	única	categoría	especializada	que	nadie	ganó
DT-F-SG	ADJ-F-SG	N-F-SG	ADJ-F-SG	WDT	PRP	VBD

$$P(\text{sent}) = \prod P(\texttt{PRP}|\texttt{DT-F-SG},\texttt{ADJ-F-SG},\texttt{N-F-SG},\texttt{ADJ-F-SG},\texttt{WDT})$$

PoS+morphology language model order

PoS+morphology language model:

La	única	categoría	especializada	que	nadie	ganó
DT-F-SG	ADJ-F-SG	N-F-SG	ADJ-F-SG	WDT	PRP	VBD

$$P(\text{sent}) = \prod P(\texttt{VBD}|\texttt{DT-F-SG},\texttt{ADJ-F-SG},\texttt{N-F-SG},\texttt{ADJ-F-SG},\texttt{WDT},\texttt{PRP})$$

PoS+morphology language model order

PoS+morphology language model:

La	única	categoría	especializada	que	nadie	ganó
DT-F-SG	ADJ-F-SG	N-F-SG	ADJ-F-SG	WDT	PRP	VBD

$$P(\text{sent}) = \prod P(\texttt{VBD}|\texttt{DT-F-SG},\texttt{ADJ-F-SG},\texttt{N-F-SG},\texttt{ADJ-F-SG},\texttt{WDT},\texttt{PRP})$$

N-gram language model:
Order 3: $P(\texttt{VBD}|\texttt{WDT},\texttt{PRP})$
Order 5: $P(\texttt{VBD}|\texttt{N-F-SG},\texttt{ADJ-F-SG},\texttt{WDT},\texttt{PRP})$
Order 7: $P(\texttt{VBD}|\texttt{DT-F-SG},\texttt{ADJ-F-SG},\texttt{N-F-SG},\texttt{ADJ-F-SG},\texttt{WDT},\texttt{PRP})$

- Croatian has a relatively free word order: is it better to use a low-order POS+morphology LM?

Constrained PoS tagging

- State-of-the-art PoS tagger for Croatian is based on a CRF classifier
- It can assign to a word a PoS+morphology tag not present in the lexicon for that word, thus increasing decoding space and potentially penalising translation quality (Ling et al., 2012)

Constrained PoS tagging

- State-of-the-art PoS tagger for Croatian is based on a CRF classifier
- It can assign to a word a PoS+morphology tag not present in the lexicon for that word, thus increasing decoding space and potentially penalising translation quality (Ling et al., 2012)

Example

Phrase table (vanilla SMT):

house	*kuća*	0.4
...	...	...

Constrained PoS tagging

- State-of-the-art PoS tagger for Croatian is based on a CRF classifier
- It can assign to a word a PoS+morphology tag not present in the lexicon for that word, thus increasing decoding space and potentially penalising translation quality (Ling et al., 2012)

Example

Phrase table (factored):

house	*kuća*|*N-F-SG-NOM*	0.09
house	*kuća*|*N-F-PL-GEN*	0.26
house	*kuća*|*NUM-CARD*	0.01
house	*kuća*|*ADJ-F-PL-LOC*	0.01
house	*kuća*|*PRON-F-PL-NOM*	0.01
house	*kuća*|*FOREIGN*	0.01
...	...	

Constrained PoS tagging

- State-of-the-art PoS tagger for Croatian is based on a CRF classifier
- It can assign to a word a PoS+morphology tag not present in the lexicon for that word, thus increasing decoding space and potentially penalising translation quality (Ling et al., 2012)

Solution

Only choose tags compatible with lexicon: reduce translation options and slightly decrease tagging accuracy

Example

Phrase table (factored):

house	*kuća*|*N-F-SG-NOM*	0.09
house	*kuća*|*N-F-PL-GEN*	0.26
...		

Results

- Results of evaluation with *newstest2012*:

PoS+morph. LM order	constrained tagging	BLEU
baseline	-	0.2356
3	no	**0.2429**
5	no	0.2408
7	no	0.2373
3	yes	**0.2458**
5	yes	**0.2432**
7	yes	0.2413

- Combination with data selection methods:
  - Select sentences that look like news from out-of-domain parallel corpora

System	BLEU
factored (from Web corpora, 1M parallel sentences)	0.2458
data selection (non-factored, 4.7M parallel sentences)	0.2576
data selection + factored (4.7M parallel sentences)	0.2700
Google Translate	0.2673

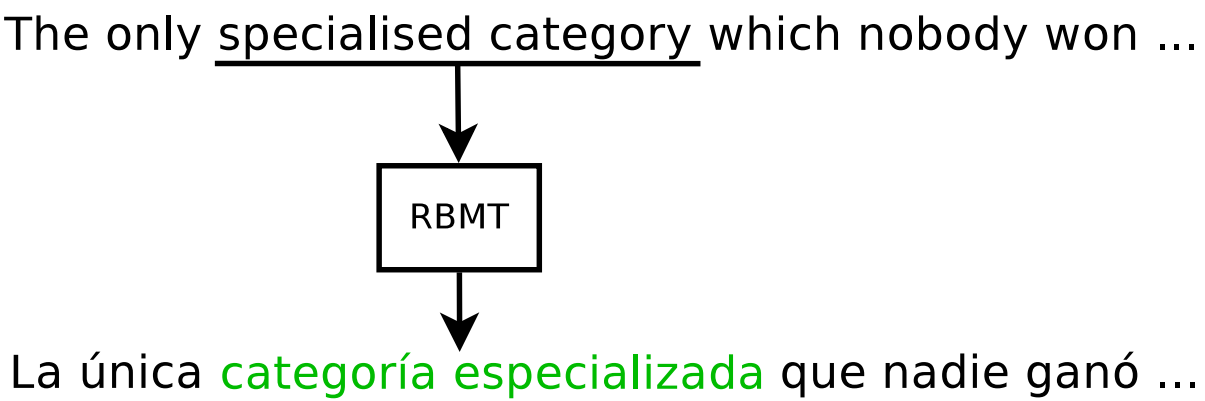
Outline

Motivation

- Our factored model setup cannot translate and generate **unseen** words in training corpus

Solution

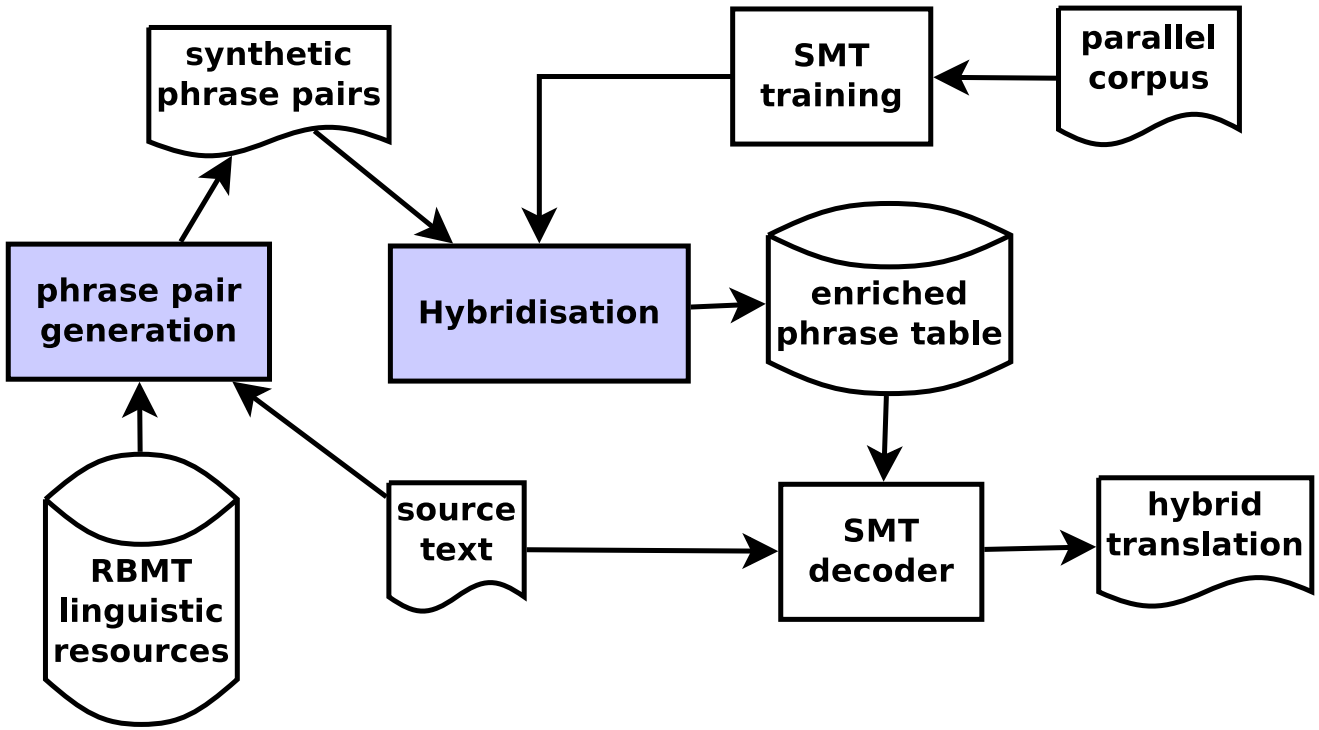
Allow the SMT system to use phrase pairs generared from shallow-transfer RBMT linguistic resources



- Developed a new method that, unlike existing *black-box* approaches in the literature, selects only **high-quality** phrases from the RBMT system

Integrating RBMT data into SMT

Method overview



- Use inner workings of RBMT translation process to generate high-quality synthetic bilingual phrases
- Integrate synthetic bilingual phrases into SMT models

Integrating RBMT data into SMT

Generation of synthetic phrase pairs

Strategy

- Generate phrase pairs for all the bilingual dictionary entries
- Segment the SL text to be translated with shallow-transfer rules

- All the linguistic information is extracted from the RBMT system without loss

Example:

SL text:

The only specialised category that nobody won ...

the `DT` *only* `ADJ` *specialised* `ADJ` *category* `N-SG` *that* `WDT` *nobody* `PRP` *won* `VBD` ...

Integrating RBMT data into SMT

Generation of synthetic phrase pairs

Strategy

- Generate phrase pairs for all the bilingual dictionary entries
- Segment the SL text to be translated with shallow-transfer rules

- All the linguistic information is extracted from the RBMT system without loss

Example:

SL text:

The only specialised category that nobody won ...

the DT *only* ADJ *specialised* ADJ *category* N-SG *that* WDT *nobody* PRP *won* VBD ...

Generated bilingual phrases:

specialised category – categoría especializada

Integrating RBMT data into SMT

Generation of synthetic phrase pairs

Strategy

- Generate phrase pairs for all the bilingual dictionary entries
- Segment the SL text to be translated with shallow-transfer rules

- All the linguistic information is extracted from the RBMT system without loss

Example:

SL text:

The only specialised category that nobody won ...

the DT *only* ADJ *specialised* ADJ *category* N-SG *that* WDT *nobody* PRP *won* VBD ...

Generated bilingual phrases:

nobody won – nadie ganó

Integrating RBMT data into SMT

Integration of synthetic phrase pairs into SMT

Multiple strategies can be followed:

- Phrase table linear interpolation
- Independent phrase tables (and decoding paths)
- Join corpus-extracted + synthetic phrase pairs, do phrase scoring and add binary feature function

Integrating RBMT data into SMT

Integration of synthetic phrase pairs into SMT

Multiple strategies can be followed:

- Phrase table linear interpolation
- Independent phrase tables (and decoding paths)
- **Join corpus-extracted + synthetic phrase pairs, do phrase scoring and add binary feature function** Best translation quality

the	*el*	0	0.5
the	*las*	0	0.2
small houses	*casas pequeñas*	0	0.7
small	*medianas*	0	0.1
specialised	*especializados*	0	1.0
specialised category	*categoría especializada*	1	1.0
...	...	...	..

Integrating RBMT data into SMT

Integration of synthetic phrase pairs into SMT

Multiple strategies can be followed:

- Phrase table linear interpolation
- **Independent phrase tables (and decoding paths)** Balance between translation quality and speed
- Join corpus-extracted + synthetic phrase pairs, do phrase scoring and add binary feature function

Original phrase table:

the	*el*	0.5
the	*las*	0.2
small houses	*casas pequeñas*	0.7
small	*medianas*	0.1
specialised	*especializados*	1.0
...	...	...

Synthetic phrase table:

specialised category	*categoría especializada*	1.0

Evaluation

Evaluation goals:

- Compare translation quality achieved by hybrid system with other approaches:
  - Pure RBMT (Apertium) and phrase-based SMT systems
  - *Black-box* hybrid approach based on statistical word alignments (Eisele et al., 2008)
- Measure impact of:
  - Size of parallel and monolingual corpora
  - Domain of test corpus: same or different from training

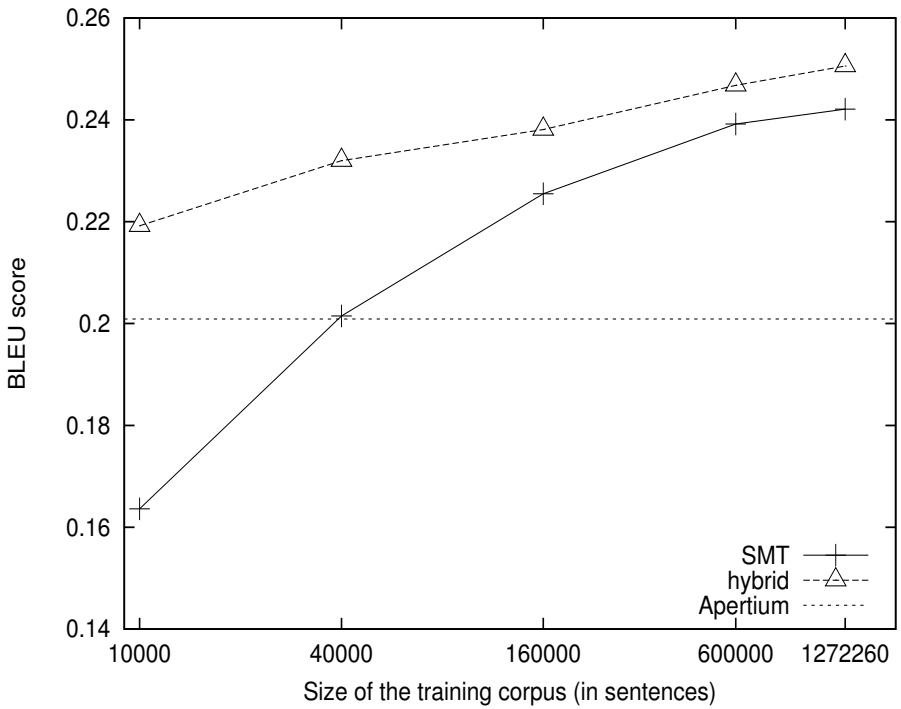
	train	out-of-domain test	TL model
English↔Spanish	*Europarl (1.2M)*	*newstest2010*	*Europarl* (+ *newscrawl*)
Breton→French	*Ofis Publik (50K)*	–	*Ofis Publik* + *Europarl*

Some results

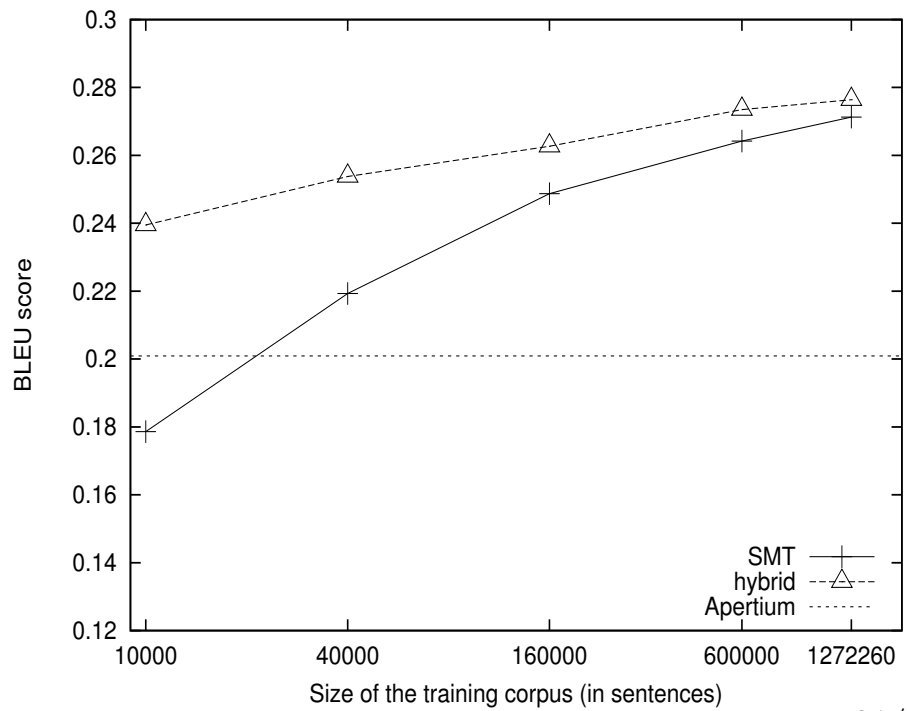
- Systematically outperforms *black-box* approach
- Outperforms pure systems when parallel corpus is very small or out-of-domain texts are translated
- Increasing the size of the language model reduces impact

Example: Spanish→English out-of-domain

TL model: *Europarl*



TL model: *Europarl* + *newscrawl* (4x bigger)



Outline

Concluding remarks

Address data sparseness in SMT with morphologically rich languages

- Generate of grammatically correct, unseen TL sequences by means of **factored translation models**
  - Order of PoS+morphology LM and tagging algorithm are important parameters
  - Reached Google Translate in English-to-Croatian by combining factored models and data selection
- Generate even unseen words by means of **integration of RBMT rules and dictionaries into SMT**
  - Highly effective for out-of-domain translation
  - Code available at:
    `https://github.com/vitaka/rule2phrase`
    Documentation coming soon!

Future research lines

- Study phrase table filtering to reduce even more number of translation alternatives with factored models
- Use Recurrent Neural Network LM (Mikolov et at., 2010) for PoS+morphology
- Tight integration of automatic rule inference (Sánchez-Cartagena et al., 2015) and hybridisation: prob. of synthetic phrases proportional to prob. of inferred rule
- Combine integration of RBMT rules and dictionaries with factored models

Thank you for your attention