An Italian to Catalan RBMT system reusing data from existing language pairs

Antonio Toral, Mireia Ginestí-Rosell, Francis Tyers

2^{nd} International Workshop on Free/Open-Source Rule-Based Machine Translation

2011/01/21

Contents

1. Introduction
2. Methodology
   - Crossdics
   - Inconsistencies
   - Coverage
3. Evaluation
   - Setting
   - Results
4. Conclusions

- Two main approaches in Machine Translation: Rule-Based and Statistical

- Two main approaches in Machine Translation: Rule-Based and Statistical
- What is needed to build a system for a new language pair?

- Two main approaches in Machine Translation: Rule-Based and Statistical
- What is needed to build a system for a new language pair?
    - RBMT: dictionaries and rules

- Two main approaches in Machine Translation: Rule-Based and Statistical
- What is needed to build a system for a new language pair?
  - RBMT: dictionaries and rules
  - SMT: parallel corpus

- Two main approaches in Machine Translation: Rule-Based and Statistical
- What is needed to build a system for a new language pair?
  - RBMT: dictionaries and rules
  - SMT: parallel corpus
- Drawbacks:

- Two main approaches in Machine Translation: Rule-Based and Statistical
- What is needed to build a system for a new language pair?
  - RBMT: dictionaries and rules
  - SMT: parallel corpus
- Drawbacks:
  - RBMT: linguistic expertise on both languages, manual construction

- Two main approaches in Machine Translation: Rule-Based and Statistical
- What is needed to build a system for a new language pair?
  - RBMT: dictionaries and rules
  - SMT: parallel corpus
- Drawbacks:
  - RBMT: linguistic expertise on both languages, manual construction
  - SMT: only applicable to language pairs with big amounts of parallel data

- This paper: build RBMT system by exploiting data from existing pairs

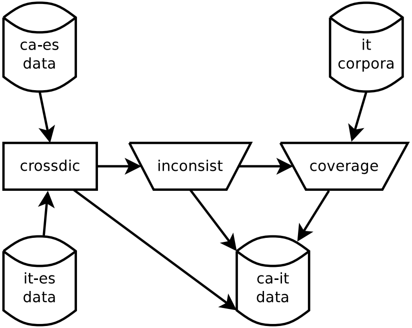
- This paper: build RBMT system by exploiting data from existing pairs
- We build an MT system for pair *a*–*b* given existing systems for pairs *a*–*c* and *b*–*c*

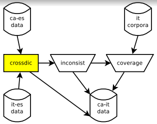
- This paper: build RBMT system by exploiting data from existing pairs
- We build an MT system for pair *a*–*b* given existing systems for pairs *a*–*c* and *b*–*c*
- Italian→Catalan from Apertium's Italian–Spanish and Catalan–Spanish

- This paper: build RBMT system by exploiting data from existing pairs
- We build an MT system for pair *a*–*b* given existing systems for pairs *a*–*c* and *b*–*c*
- Italian→Catalan from Apertium's Italian–Spanish and Catalan–Spanish
- Motivation:

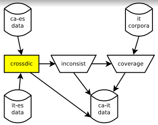
- This paper: build RBMT system by exploiting data from existing pairs
- We build an MT system for pair *a*–*b* given existing systems for pairs *a*–*c* and *b*–*c*
- Italian→Catalan from Apertium's Italian–Spanish and Catalan–Spanish
- Motivation:
  - RBMT competitive and useful for languages without parallel corpora

- This paper: build RBMT system by exploiting data from existing pairs
- We build an MT system for pair *a–b* given existing systems for pairs *a–c* and *b–c*
- Italian→Catalan from Apertium's Italian–Spanish and Catalan–Spanish
- Motivation:
  - RBMT competitive and useful for languages without parallel corpora
  - Reusing data from similar pairs significantly reduces the amount of work

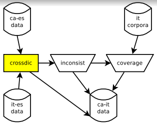

ca-es data
it corpora
crossdic
inconsist
coverage
it-es data
ca-it data



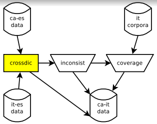
- Input dictionaries:



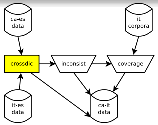
- Input dictionaries:
  - es–it: mono es 11k, mono it 10k, bi 12k



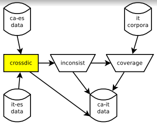
- Input dictionaries:
  - es–it: mono es 11k, mono it 10k, bi 12k
  - es–ca: mono es 44k, mono ca 40k, bi 51k



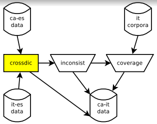
- Input dictionaries:
  - es–it: mono es 11k, mono it 10k, bi 12k
  - es–ca: mono es 44k, mono ca 40k, bi 51k
- Output dictionaries:



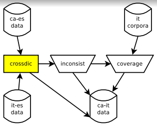
- Input dictionaries:
  - es–it: mono es 11k, mono it 10k, bi 12k
  - es–ca: mono es 44k, mono ca 40k, bi 51k
- Output dictionaries:
  - it–ca: mono it 7k, mono ca 8k, bi 9k



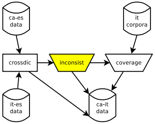
- Input dictionaries:
  - es–it: mono es 11k, mono it 10k, bi 12k
  - es–ca: mono es 44k, mono ca 40k, bi 51k
- Output dictionaries:
  - it–ca: mono it 7k, mono ca 8k, bi 9k
- Other linguistic data:



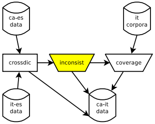
- Input dictionaries:
  - es–it: mono es 11k, mono it 10k, bi 12k
  - es–ca: mono es 44k, mono ca 40k, bi 51k
- Output dictionaries:
  - it–ca: mono it 7k, mono ca 8k, bi 9k
- Other linguistic data:
  - it tagger and disambiguation probabilities taken from it–es



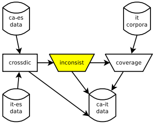
- Input dictionaries:
  - es–it: mono es 11k, mono it 10k, bi 12k
  - es–ca: mono es 44k, mono ca 40k, bi 51k
- Output dictionaries:
  - it–ca: mono it 7k, mono ca 8k, bi 9k
- Other linguistic data:
  - it tagger and disambiguation probabilities taken from it–es
  - transfer rules: 35 taken from oc–ca (mainly noun phrases) + 9 manually created (verbs and clitic pronouns)


ca-es data
it corpora
crossdic
inconsist
coverage
it-es data
ca-it data

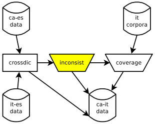
- Reasons



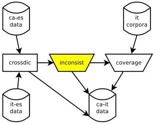
- Reasons
  - Differences of gender and number (it–ca)



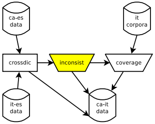
- Reasons
  - Differences of gender and number (it–ca)
  - Different ways of categorising lemmas and morphological features (es–it and es–ca)



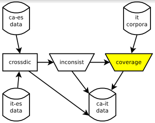
- Reasons
  - Differences of gender and number (it–ca)
  - Different ways of categorising lemmas and morphological features (es–it and es–ca)
- Solutions



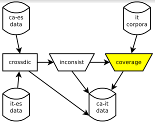
- Reasons
  - Differences of gender and number (it–ca)
  - Different ways of categorising lemmas and morphological features (es–it and es–ca)
- Solutions
  - Manually solve inconsistencies (identified automatically), 0.5 Person Months



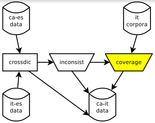
- Reasons
  - Differences of gender and number (it–ca)
  - Different ways of categorising lemmas and morphological features (es–it and es–ca)
- Solutions
  - Manually solve inconsistencies (identified automatically), 0.5 Person Months
  - Substitute derived ca mono dictionary (8k) for that in es–ca (40k)



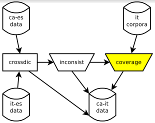
- Coverage calculated on two Italian corpora: Europarl and Wikipedia



- Coverage calculated on two Italian corpora: Europarl and Wikipedia
- 155 most frequent unknown words added to the system



- Coverage calculated on two Italian corpora: Europarl and Wikipedia
- 155 most frequent unknown words added to the system



- Coverage calculated on two Italian corpora: Europarl and Wikipedia
- 155 most frequent unknown words added to the system

	Europarl	Wikipedida
Number of words	46,569,602	241,563,615
Initial coverage	86.4%	75.5%
Final coverage	88.9%	79.4%

- Systems

- Systems
    - Apertium (the it→ca system)

- Systems
  - Apertium (the it→ca system)
  - Apertium-i (indirect translation using it→es and es→ca)

- Systems
  - Apertium (the it→ca system)
  - Apertium-i (indirect translation using it→es and es→ca)
  - Google Translate

- Systems
  - Apertium (the it→ca system)
  - Apertium-i (indirect translation using it→es and es→ca)
  - Google Translate
- Test set: 1k sentences from KDE4 (OPUS project)

- Systems
  - Apertium (the it→ca system)
  - Apertium-i (indirect translation using it→es and es→ca)
  - Google Translate
- Test set: 1k sentences from KDE4 (OPUS project)
- Metrics: TER, GTM, BLEU, NIST

Metric	Apertium	Apertium-i	Google
TER	**0.5703**	0.6118	0.6785
GTM	**0.5162**	0.4712	0.41637
BLEU	0.2290	0.1492	**0.2459**
NIST	5.6567	4.4753	**6.1071**

- GTM and TER: Apertium $>$ Apertium-i $>$ Google

Metric	Apertium	Apertium-i	Google
TER	**0.5703**	0.6118	0.6785
GTM	**0.5162**	0.4712	0.41637
BLEU	0.2290	0.1492	**0.2459**
NIST	5.6567	4.4753	**6.1071**

- GTM and TER: Apertium > Apertium-i > Google
- BLEU*: Google ≃ Apertium > Apertium-i

Metric	Apertium	Apertium-i	Google
TER	**0.5703**	0.6118	0.6785
GTM	**0.5162**	0.4712	0.41637
BLEU	0.2290	0.1492	**0.2459**
NIST	5.6567	4.4753	**6.1071**

- GTM and TER: Apertium > Apertium-i > Google
- BLEU*: Google $\simeq$ Apertium > Apertium-i
- NIST*: Google > Apertium > Apertium-i

- it→ca RBMT system derived from es–it and es–ca systems

- it→ca RBMT system derived from es–it and es–ca systems
- Limited amount of manual work: correct inconsistencies, augment coverage and add some transfer rules

- it→ca RBMT system derived from es–it and es–ca systems
- Limited amount of manual work: correct inconsistencies, augment coverage and add some transfer rules
- Evaluated against RBMT indirect system and SMT system, yielding significant improvements over both

- it→ca RBMT system derived from es–it and es–ca systems
- Limited amount of manual work: correct inconsistencies, augment coverage and add some transfer rules
- Evaluated against RBMT indirect system and SMT system, yielding significant improvements over both
- System released as apertium-ca-it-0.1.0

Thanks! Questions?