MT4ALL

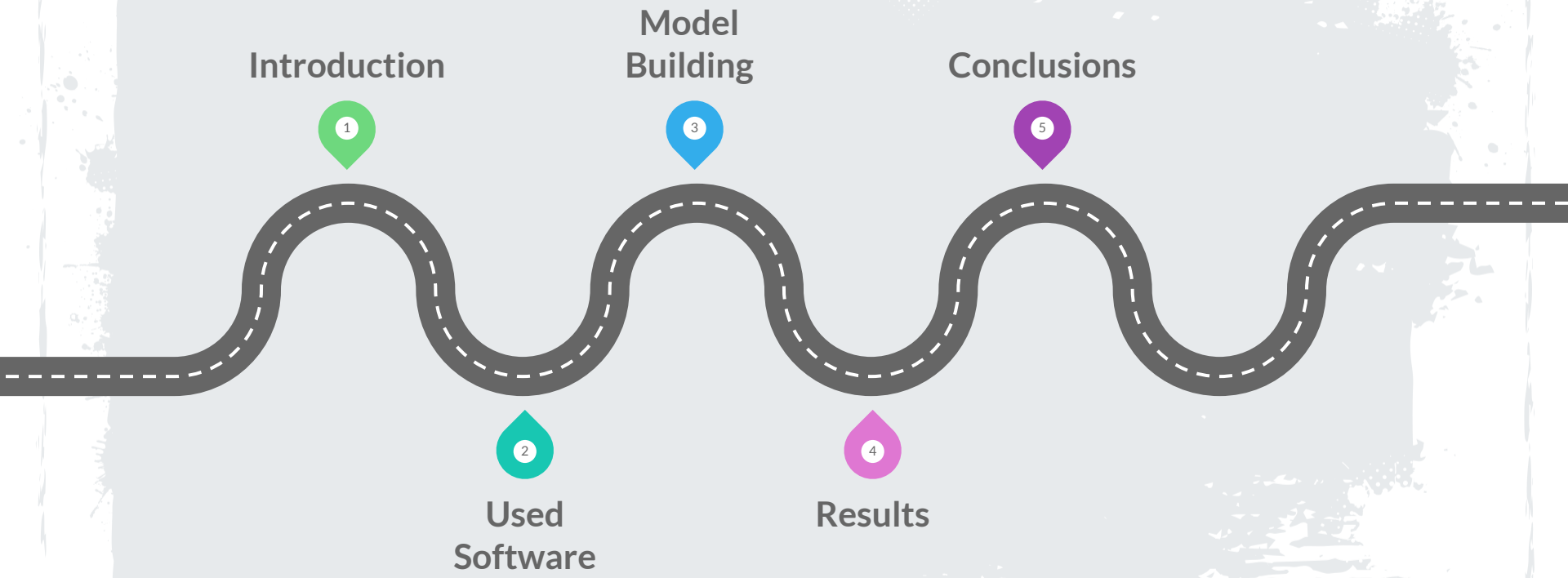
Universidad del País Vasco
Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

Technical Approach and Results

Iakes Goenaga

Outline


Introduction
1
Used Software
2
Model Building
3
Results
4
Conclusions
5

1. Introduction

Introduction (I)

Main objective:

Generate MT related bilingual resources for those language pairs lacking sufficient parallel corpora.

Introduction (I)

Main objective:

Generate MT related bilingual resources for those language pairs lacking sufficient parallel corpora.

- Unsupervised approach

Introduction (I)

Main objective:

Generate MT related bilingual resources for those language pairs lacking sufficient parallel corpora.

- Unsupervised approach
- Supervised approach

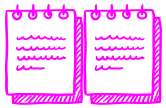
Introduction (I)

Main objective:

Generate MT related bilingual resources for those language pairs lacking sufficient parallel corpora.

- × Unsupervised approach
- × Supervised approach

Introduction (II)

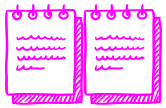


Supervised models:

Cons:

Pros:

Introduction (II)



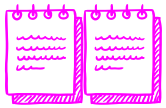
Supervised models:

Cons:

- × Parallel corpus

Pros:

Introduction (II)



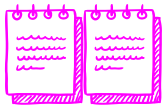
Supervised models:

Cons:

- × Parallel corpus
  - × Time 
  - × Cost 

Pros:

Introduction (II)



Supervised models:

Cons:

- Parallel corpus
  - Time
  - Cost

Pros:

- Training time
- Results

Introduction (III)

Unsupervised models:

Cons:

Pros:

Introduction (III)

Unsupervised models:

Cons:

- × Training time
- × Results (parallel corpora exists)

Pros:

Introduction (III)

Unsupervised models:

Cons:

- × Training time
- × Results (parallel corpora exists)

Pros:

- × Cost
- × Results (parallel corpora doesn't exist)

Introduction (III)

Unsupervised models:

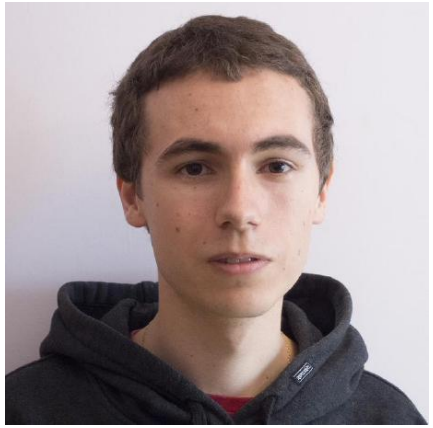
Cons:

- × Training time
- × Results (parallel corpora exists)

Pros:

- × Cost
- × Results (parallel corpora doesn't exist)

Introduction (IIII)



Mikel Artetxe:

An Effective Approach to Unsupervised Machine Translation (2019)


Why?

Introduction (V)

3 main reasons:

- Promising results
- Only monolingual corpora needed
- Good results on high resourced languages
  - Experiment on low resourced languages

2. Used Software

Used Software (I)

Monoses

https://github.com/artetxem/monoses

Used Software (I)

Monoses
https://github.com/artetxem/monoses

- Hybrid approach:
  - Unsupervised MT for initialization
  - Dual NMT model trained through iterative-backtranslation

Used Software (I)

Monoses
https://github.com/artetxem/monoses

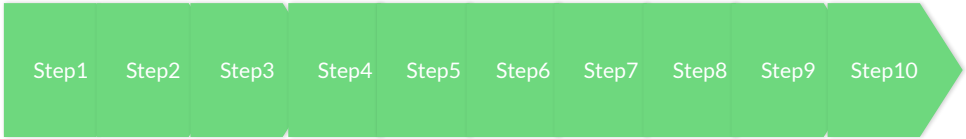
- Hybrid approach:
  - Unsupervised MT for initialization
  - Dual NMT model trained through iterative-backtranslation

- Interesting results:
  - WMT-14
    - FR-EN: 33.5 | EN-FR: 36.2
    - DE-EN: 27.0 | EN-DE: 22.5

Used Software (II)

Monoses
https://github.com/artetxem/monoses

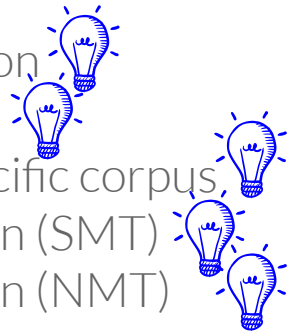
- The process is divided in 10 steps



Used Software (II)

Monoses
https://github.com/artetxem/monoses

- Added features:
  - Domain-specific corpus addition
  - BPE application from begining
  - Oversampling for domain-specific corpus
  - Continue in a previous iteration (SMT)
  - Continue in a previous iteration (NMT)



3. Model Building

Model Building (I)

- 12 models:
  - 6 domains:
    - General domain
    - Newswire domain
    - Financial domain
    - Legal domain
    - Biomedical domain
    - Customer support domain

Model Building (II)

- 11 languages:
  - English
  - Basque
  - German
  - Finnish
  - Latvian
  - Georgian
  - Kazakh
  - Ukrainian
  - Catalan
  - Norwegian
  - Spanish

Model Building (III)

- General domain:
  - English-Catalan

Model Building (III)

- General domain:
  - English-Catalan

- English:
  - 149M sentences (OSCAR + Newscrawl)
- Catalan:
  - 72M sentences (MT4All)

Model Building (IV)

- Newswire domain:
  - English-Basque

Model Building (IV)

- Newswire domain:
  - English-Basque

- English:
  - 149M sentences (OSCAR + Newscrawl)
- Basque:
  - 18M sentences (OSCAR + MT4All + Elhuyar)

Model Building (V)

- Biomedical domain:
  - English-Spanish

Model Building (V)

- Biomedical domain:
  - English-Spanish

- English:
  - 75 M sentences (MT4All)
- Spanish:
  - 41M sentences (MT4All)

Model Building (VI)

- Customer support domain:
  - English-Spanish
  - English-German
  - English-Norwegian

Model Building (VI)

- English:
  - 149M sentences (OSCAR + Newscrawl)
  - 6M sentences (MT4All)
- Spanish:
  - 131M sentences (OSCAR)
  - 998K sentences (MT4All)
- German:
  - 157M sentences (OSCAR)
  - 5M sentences (MT4All)
- Norwegian:
  - 31M sentences (OSCAR)
  - 3M sentences (MT4All)

Model Building (VI)

- English:
  - 149M sentences (OSCAR + Newscrawl)
  - 6M sentences (MT4All) x 4 (Oversampling) 
- Spanish:
  - 131M sentences (OSCAR)
  - 998K sentences (MT4All) x 6 (Oversampling) 
- German:
  - 157M sentences (OSCAR)
  - 5M sentences (MT4All) x 4 (Oversampling) 
- Norwegian:
  - 31M sentences (OSCAR)
  - 3M sentences (MT4All) x 6 (Oversampling) 

Model Building (VII)

- Legal domain:
  - English-Georgian
  - English-Kazakh
  - English-Ukrainian

Model Building (VII)

- English:
  - 149M sentences (OSCAR + Newscrawl)
  - 147K sentences (MT4All)
- Georgian:
  - 5M sentences (OSCAR)
  - 203K sentences (MT4All)
- Kazakh:
  - 9M sentences (OSCAR)
  - 124K sentences (MT4All)
- Ukrainian:
  - 84M sentences (OSCAR)
  - 7M sentences (MT4All)

Model Building (VII)

- English:
  - 149M sentences (OSCAR + Newscrawl)
  - 147K sentences (MT4All) x 8 (Oversampling) 
- Georgian:
  - 5M sentences (OSCAR)
  - 203K sentences (MT4All) x 8 (Oversampling) 
- Kazakh:
  - 9M sentences (OSCAR)
  - 124K sentences (MT4All) x 8 (Oversampling) 
- Ukrainian:
  - 84M sentences (OSCAR)
  - 7M sentences (MT4All) x 4 (Oversampling) 

Model Building (VIII)

- Financial domain:
  - English-Latvian
  - English-Finnish
  - English-Norwegian

Model Building (VIII)

- English:
  - 149M sentences (OSCAR + Newscrawl)
  - 3M sentences (MT4All)
- Latvian:
  - 14M sentences (OSCAR)
  - 480K sentences (MT4All)
- Finnish:
  - 112M sentences (OSCAR)
  - 3M sentences (MT4All)
- Norwegian:
  - 31M sentences (OSCAR)
  - 5M sentences (MT4All)

Model Building (VIII)

- English:
  - 149M sentences (OSCAR + Newscrawl)
  - 3M sentences (MT4All) x 6 (Oversampling)
- Latvian:
  - 14M sentences (OSCAR)
  - 480K sentences (MT4All) x 8 (Oversampling)
- Finnish:
  - 112M sentences (OSCAR)
  - 3M sentences (MT4All) x 6 (Oversampling)
- Norwegian:
  - 31M sentences (OSCAR)
  - 5M sentences (MT4All) x 4 (Oversampling)

Model Building (IX)

- 4 GPUs used to train each model:
  - NVIDIA V100 (Volta) 16GB
- 80 CPUs:
  - IBM Power9 8335-GTH
- 10 days to train each model (average):
  - 7 days for SMT model
  - 3 days for NMT model

4. Results

Results

General domain
EN-CA
30.8
CA-EN
33.4

Sacrebleu
https://github.com/mjpost/sacrebleu

CA = Catalan

Results

General domain
EN-CA
30.8 
CA-EN
33.4 

Sacrebleu
https://github.com/mjpost/sacrebleu

CA = Catalan

Results

Newswire domain
EN-EU
5.1
EU-EN
12.1

Sacrebleu
https://github.com/mjpost/sacrebleu

EU = Basque

Results

Newswire domain
EN-EU
5.1
EU-EN
12.1

Sacrebleu
https://github.com/mjpost/sacrebleu

EU = Basque

Results

Newswire domain	
EN-EU	EN-EU (BPE)
5.1	9.0
EU-EN	EU-EN (BPE)
12.1	16.0

Sacrebleu
https://github.com/mjpost/sacrebleu

EU = Basque

Results

Biomedical domain
EN-ES
41.6
ES-EN
39.7

ES = Spanish

Results

Biomedical domain
EN-ES
41.6
ES-EN
39.7

ES = Spanish

Results

Customer support domain		
EN-ES	EN-DE	EN-NO
30.3	30.6	28.9
ES-EN	DE-EN	NO-EN
33.3	35.2	31.4

ES = Spanish
DE = German
NO = Norwegian

Results

Customer support domain		
EN-ES	EN-DE	EN-NO
30.3	30.6	28.9
ES-EN	DE-EN	NO-EN
33.3	35.2	31.4

ES = Spanish
DE = German
NO = Norwegian

Results

Financial domain		
EN-LV	EN-FI	EN-NO
24.7	11.1	27.5
LV-EN	FI-EN	NO-EN
15.1	8.8	23.4

LV = Latvian
FI = Finnish
NO = Norwegian

Results

Financial domain			
EN-LV	EN-FI	EN-FI (BPE)	EN-NO
24.7	11.1	17.5 	27.5
LV-EN	FI-EN	FI-EN (BPE)	NO-EN
15.1	8.8	21.6 	23.4

LV = Latvian
FI = Finnish
NO = Norwegian

Results

Legal domain		
EN-KA	EN-KK	EN-UK
12.0	6.4	14.2
KA-ES	KK-EN	UK-EN
18.6	7.7	15.4

KA = Georgian
KK = Kazakh
UK = Ukrainian

6. Conclusions

Conclusions

- Domain corpus:
  - Big corpus
    - Good results
  - Small corpus
    - Mix with a general domain corpus

Conclusions

- Domain corpus:
  - Big corpus
    - Good results 
  - Small corpus
    - Mix with a general domain corpus

Conclusions

- Domain corpus:
  - Big corpus
    - Good results
  - Small corpus
    - Mix with a general domain corpus

Conclusions

- Domain corpus:
  - Big corpus
    - Good results 
  - Small corpus
    - Mix with a general domain corpus 
- Word segmentation (BPE) can help in morphologically rich languages
  - Basque and Finnish

Conclusions

- Domain corpus:
  - Big corpus
    - Good results 
  - Small corpus
    - Mix with a general domain corpus 
- Word segmentation (BPE) can help in morphologically rich languages
  - Basque and Finnish 

Unanswered Questions

- Is the application of BPE beneficial for morphologically not rich languages?

Unanswered Questions

- Is the application of BPE beneficial for morphologically not rich languages?
- What size of domain-specific corpus can we consider large enough to train without mixing it with a general domain corpus?

Unanswered Questions

- Is the application of BPE beneficial for morphologically not rich languages?
- What size of domain-specific corpus can we consider large enough to train without mixing it with a general domain corpus?
- When domain-specific corpus is big
  - Is it beneficial to mix it with a general domain corpus?

Thanks!

Any questions?

You can find me at:

iakes.goenaga@ehu.eus