Overview of TweetMT

A Shared Task on Machine Translation of Tweets at SEPLN 2015

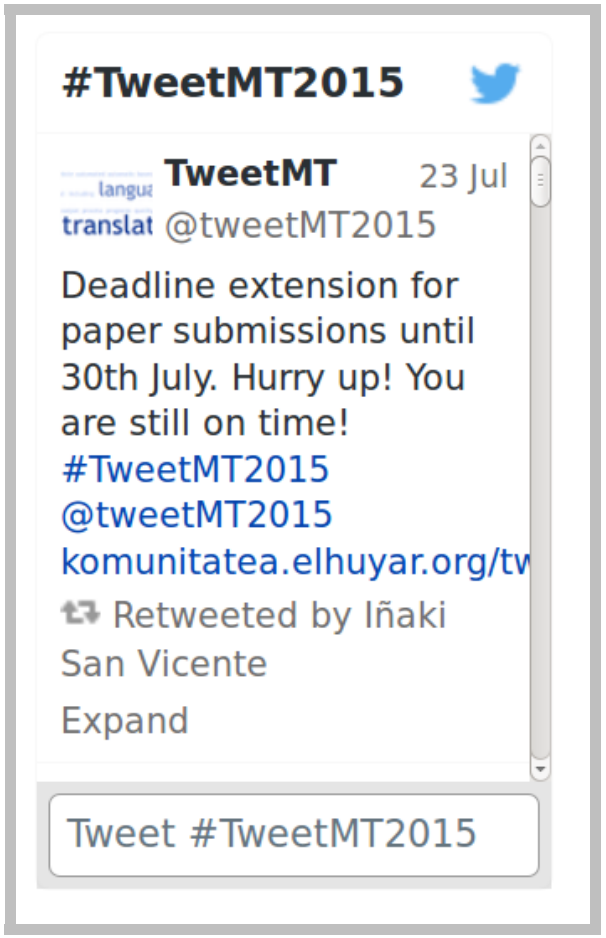
Iñaki Alegria, Nora Aranberri, **Cristina España-Bonet**,
Pablo Gamallo, Hugo Gonçalo Oliveira, Eva Martínez,
Iñaki San Vicente, Antonio Toral, Arkaitz Zubiaga

Universitat Politècnica de Catalunya, TALP Research Center

Alacant, 15th September, 2015

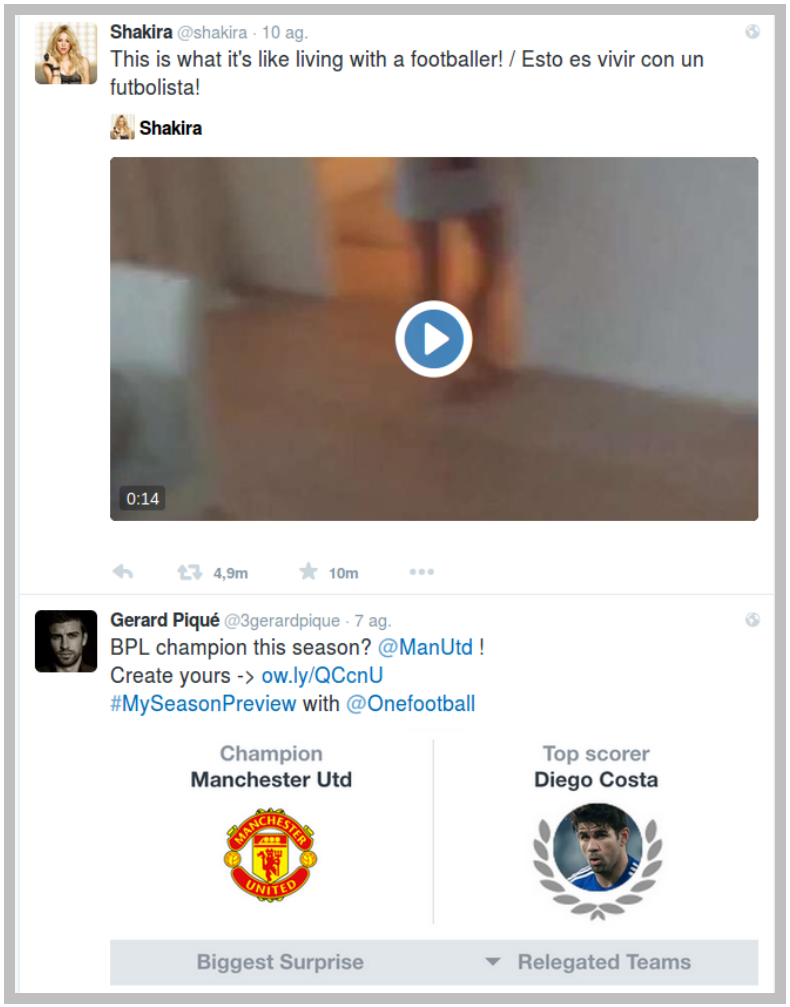
Motivation

Microblogging



Motivation

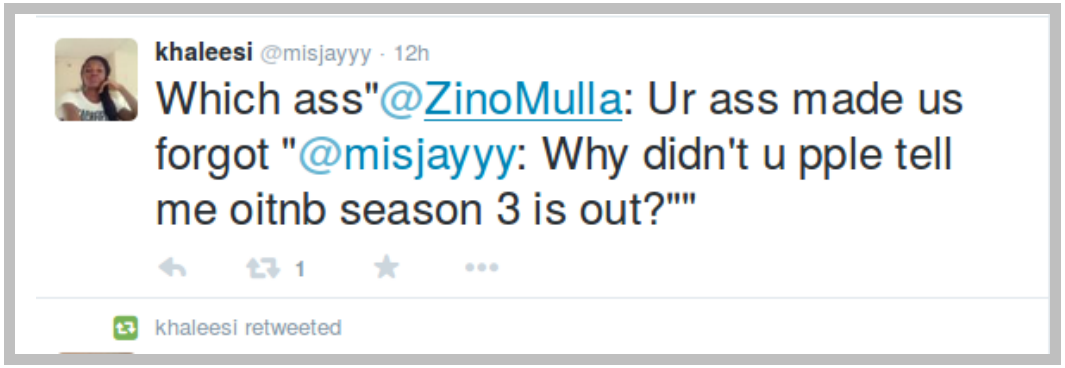
Why a specific shared task?



Text includes multiple **languages**, **hashtags**, **user** mentions, **URLs**, and all within a **length** limit

Motivation

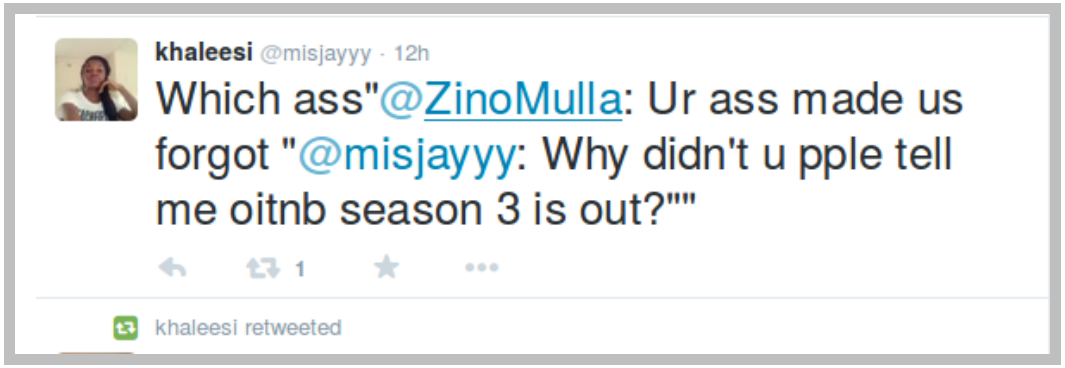
Why a specific shared task?



Familiar language,
slang,
spelling errors...

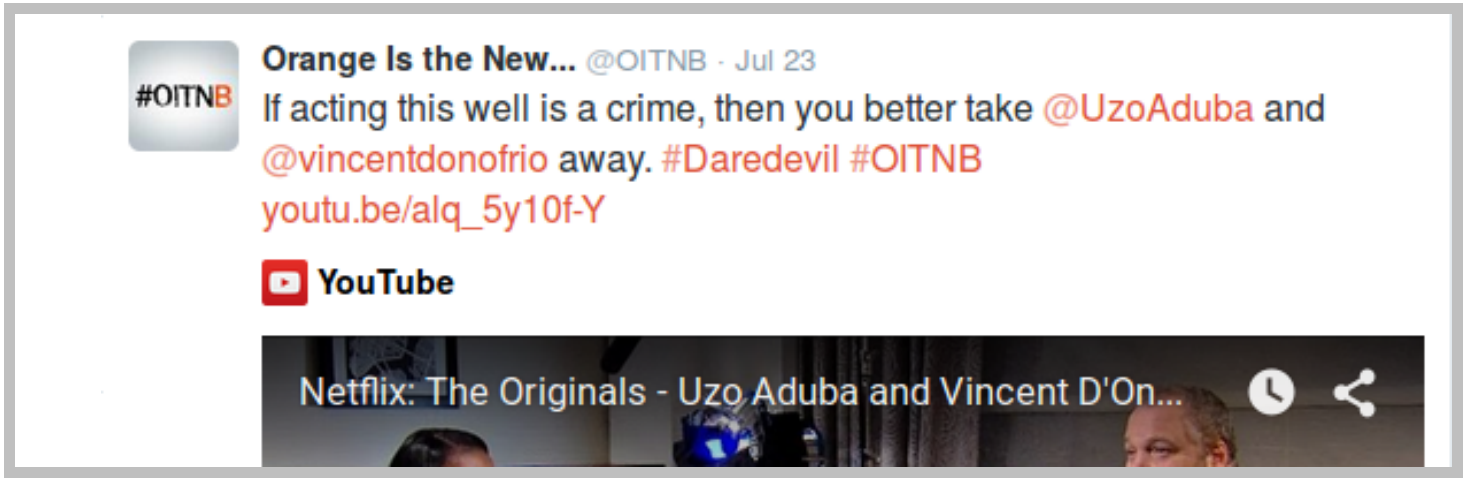
Motivation

Why a specific shared task?



Familiar language,
slang,
spelling errors...

But!



Motivation

Framework: a follow-up workshop

(Formal) Tweet Translation Task

TweetMT @SEPLN 2015 [Alegria et al., 2015]

Motivation

Framework: a follow-up workshop

3. (Formal) Tweet Translation Task

TweetMT @SEPLN 2015 [Alegria et al., 2015]

Motivation

Framework: a follow-up workshop

3. (Formal) Tweet Translation Task

TweetMT @SEPLN 2015 [Alegria et al., 2015]

2. Tweet Language Identification

TweetLID @SEPLN 2014 [Zubiaga et al., 2014]

Motivation

Framework: a follow-up workshop

3. (Formal) Tweet Translation Task

TweetMT @SEPLN 2015 [Alegria et al., 2015]

2. Tweet Language Identification

TweetLID @SEPLN 2014 [Zubiaga et al., 2014]

1. Lexical Normalisation of Tweets

TweetNorm @SEPLN 2013 [Alegria et al., 2013]

Motivation

TweetMT: Centres Involved

CISUC

elhuyar

CiTIUS
Centro Singular de Investigación en Tecnoloxías da Información

eman ta zabal zazu
universidad del país vasco
euskal herriko unibertsitatea
ixa

DCU

TALP
UPC

THE UNIVERSITY OF WARWICK

Motivation

TweetMT: Projects Involved

TACARDI: Traducción automática en contexto y aumentada con recursos dinámicos de internet

CELTIC: Conocimiento Estratégico Liderado por Tecnologías para la Inteligencia Competitiva

Abu-MaTran: Automatic building of Machine Translation

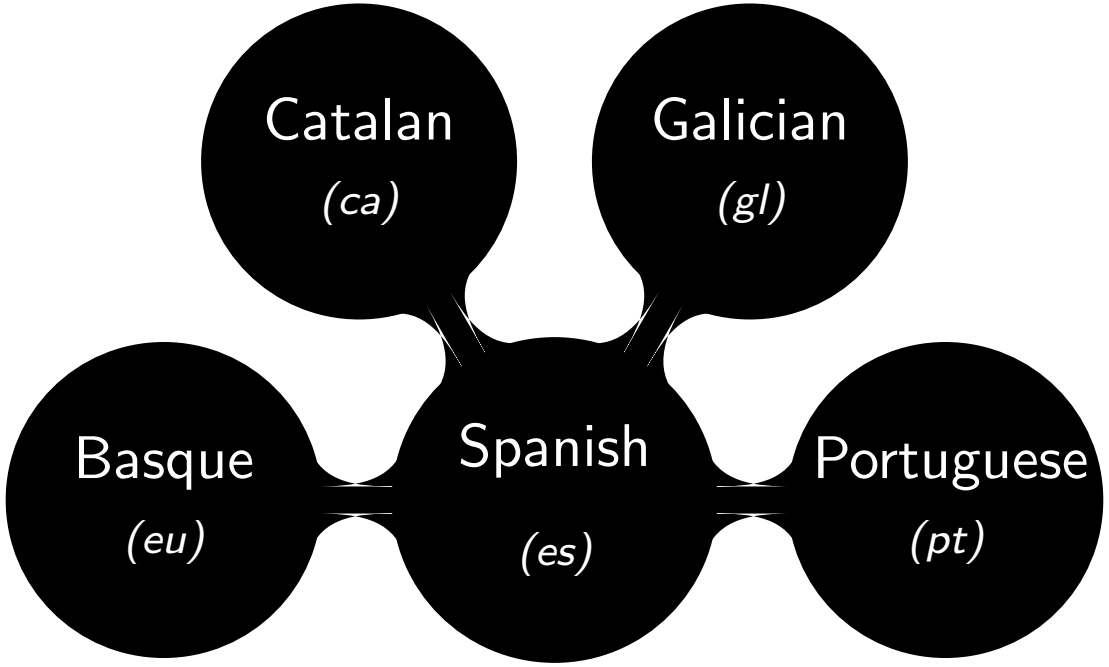
Motivation

Overview

The Shared Task

Objectives

Machine translation of tweets
between languages of the **Iberian Peninsula**



The Shared Task

Challenges

For the organisers

- Defining the problem, the task and its evaluation
- Creating from scratch the datasets

For the participants

- Developing an MT system for the given language pairs
- Dealing with peculiarities of tweets

The Shared Task

Challenges

For the organisers

- Defining the problem, the task and its evaluation
- Creating from scratch the datasets

The Shared Task

Challenges

For the organisers

- Defining the problem, the task and its **evaluation**
  - Formal tweets
  - Automatic evaluation with some normalisation
  - Discussion on informal tweet MT and evaluation

The Shared Task

Challenges

For the organisers

- Defining the problem, the task and its **evaluation**
  - Formal tweets
  - Automatic evaluation with some normalisation
  - Discussion on informal tweet MT and evaluation
- Creating from scratch the **datasets**
  - Creation from multilingual accounts
  - Creation using CrowdFlower

Building the Datasets

Overview

Building the Datasets

The problem

- No available parallel corpora
  (needed *at least* for evaluation)
- Extract parallel tweets from **parallel accounts**
  (*ca-es*, *eu-es*)

Building the Datasets

The problem

- No available parallel corpora
  (needed *at least* for evaluation)
- Extract parallel tweets from **parallel accounts**
  (*ca-es*, *eu-es*)
  - No available parallel accounts
  - Create parallel tweets via **crowdsourcing**
    (*ga-es*, *pt-es*)

Building the Datasets

Creation from Multilingual Accounts

Step 1. Identifying the accounts

Some organizations and famous personalities tend to post messages in various languages

- in a single account or,
- in multiple accounts.

Building the Datasets

Creation from Multilingual Accounts

Step 1. Identifying the accounts

Some organizations and famous personalities tend to post messages in various languages

- in a single account or,
- in multiple accounts.

We have **manually** selected

- 23 accounts (from 16 authors) for the *eu-es* pair,
- and 19 accounts (from 14 authors) for the *ca-es* pair.

Building the Datasets

Creation from Multilingual Accounts

Step 2. Data collection

Between November 2013 and March 2015

- 75,000 tweets for the *eu-es* pair
- 51,000 tweets for the *ca-es* pair

Building the Datasets

Creation from Multilingual Accounts

Step 3. Alignment

1. In-house **language identification**

Building the Datasets

Creation from Multilingual Accounts

Step 3. Alignment

1. In-house **language identification**
2. Pipeline of heuristics for the **alignment**
   - Publication date within a range
   - Overlap of hashtag and user mentions
   - Longest common subsequence ratio

Building the Datasets

Creation from Multilingual Accounts

Step 3. Alignment

1. In-house **language identification**
2. Pipeline of heuristics for the **alignment**
   - Publication date within a range
     $\Rightarrow$ cleaning
   - Overlap of hashtag and user mentions
     $\Rightarrow$ gross alignment
   - Longest common subsequence ratio
     $\Rightarrow$ slight polishment

Building the Datasets

Creation from Multilingual Accounts

Step 3. Alignment

1. In-house **language identification**
2. Pipeline of heuristics for the **alignment**
   - Publication date within a range
     $\Rightarrow$ cleaning
   - Overlap of hashtag and user mentions
     $\Rightarrow$ gross alignment
   - Longest common subsequence ratio
     $\Rightarrow$ slight polishment
3. Manual **correction**

Building the Datasets

Creation Using Crowdsourcing

Input test to be translated to Portuguese and Galician

- 2,552 Spanish tweets from the *ca-es* and *eu-es* sets
- Working sets of 10 tweets

Instructions for translators

- Not to translate user mentions and URLs
- Translate hashtags only if considered natural in the target language

Building the Datasets

Creation Using Crowdsourcing

CrowdFlower used as a crowdsourcing platform

Platform configuration:

- Only users from Spain for Galician and Portugal/Brazil for Portuguese were allowed
- Only top performance users allowed for Portuguese, had to be lowered for Galician
- Only verified Portuguese speakers
- Minimum time of 150 seconds/10 tweets considered

Building the Datasets

CrowdFlower

Traduz 10 Tweets Do Castelhano Para O PortuguéS

Instructions

Cada trabalho (job) contém 10 tweets em castelhano. A tua tarefa consiste em traduzi-los para o português. É importante teres em conta as seguintes instruções já que se não as cumprires serás automaticamente removido/a durante a execução da tarefa:

- Os tweets devem ser traduzidos manualmente. O uso de programas de tradução automática NÃO é permitido.
- As menções a utilizadores (palavras que começam por @) e as URLS (palavras que começam por http) NÃO devem ser traduzidas
- Os hashtags (palavras que começam por #) traduzir-se-ão ou não dependendo se em Português é mais natural usar o hashtag original ou traduzido.

Aqui tens uma tradução de exemplo:

- #EnDirecto El #eclipse2015 en las islas Feroe... y en el especial de @elmundovideos http://t.co/PNglvAvYNo
- #EmDirecto O #eclipse2015 nas ilhas Faroe... e no especial de @elmundovideos http://t.co/PNglvAvYNo

. @Berreibarkultur y @Amara_Bai seleccionados pararealizar un #AuzoLab #Donostia2016 @AuzolabDSS2016

Tradução em Português

La gastronomía estuvo muy presente durante los 3 días de festival. https://t.co/BAfUEpKH12 @cocinandocon_ @bculinary @labretxa @petritegi

Tradução em Português

¿Quieres crear accesorios para tu bici en una impresora 3D? Hoy en @Hirikilabs http://t.co/WYlGvt5Mfj #Hirikikas @Tabakalera

Tradução em Português

Building the Datasets

Summary

Dataset	Tweets	Authors	Tokens	URL	@user
eu-es$_{\text{dev}}$	4,000	4	181K	2,622	1,569
ca-es$_{\text{dev}}$	4,000	2	161K	3,280	823
eu-es$_{\text{test}}$	2,000	16	37K	1,556	673
es-eu$_{\text{test}}$	2,000	16	43K	1,535	692
ca-es$_{\text{test}}$	2,000	14	45K	1,590	417
es-ca$_{\text{test}}$	2,000	14	46K	1,567	502
gl-es$_{\text{test}}$	388	*10*	7K	274	134
es-gl$_{\text{test}}$	388	*10*	7K	291	159
pt-es$_{\text{test}}$	1,250	*40*	19K	674	349
es-pt$_{\text{test}}$	1,250	*40*	21K	919	583

Systems at Competition

Overview

Systems at Competition

Participation

- **5** registered **participants**
  Only 3 submitted: DCU, EHU & UPC
- **10 submissions** from 3 participants
  Language pairs: *ca-es*, *eu-es* & *pt-es*
  (none for *gl-es*)
- **24** translation **systems**
  9 for *ca-es*, 9 for *eu-es* & 6 for *pt-es*

Systems at Competition

Participation by Language Pair

ca-es
DCU,UPC
9 sys

eu-es
DCU,EHU
9 sys

pt-es
DCU
6 sys

Systems at Competition

Participation by Team

DCU

ca-es,eu-es,pt-es

18 sys

EHU

eu-es

3 sys

UPC

ca-es

3 sys

Systems at Competition

Systems I: DCU

	System combination (best in dev used for test)
DCU1	*es* $\leftrightarrow$ *ca*: Moses and Apertium
	es $\rightarrow$ *eu*: Moses, cdec and Apertium
	eu $\rightarrow$ *es*: cdec
	es $\leftrightarrow$ *pt*: Moses
DCU2	*es* $\rightarrow$ *ca*: Moses
	ca $\rightarrow$ *es*, *eu* $\rightarrow$ *es*: Moses, cdec and Apertium
	es $\rightarrow$ *eu*: Moses, cdec, ParFDA, Matxin and Morph
	es $\leftrightarrow$ *pt*: Moses and cdec
DCU3	*es* $\rightarrow$ *ca*, *es* $\leftrightarrow$ *pt*: Moses, cdec and Apertium
	ca $\rightarrow$ *es*: Moses, ParFDA and Apertium
	es $\rightarrow$ *eu*: Moses, cdec, Matxin and Morph
	eu $\rightarrow$ *es*: Moses, cdec, Apertium and Morph

Systems at Competition

Systems II: UPC, EHU

	Main Engine	Distinctive features
EHU1	SMT	Specific language model and pre- and post-processing for tweets
EHU2	RBMT	Adaptation to tweets (mainly hashtags)

Systems at Competition

Systems II: UPC, EHU

	Main Engine	Distinctive features
EHU1	SMT	Specific language model and pre- and post-processing for tweets
EHU2	RBMT	Adaptation to tweets (mainly hashtags)

	Main Engine	Distinctive features
UPC1	SMT	In-genre Moses system
UPC2	SMT	Document-level system (Docent) and semantic models

Evaluation

Overview

Evaluation

Framework I

Heterogeneous set of automatic metrics

– **Lexical metrics**

- **PER, TER, WER**: Subset of metrics based on edit distances
- **BLEU, NIST, ROUGE**: Based on *n*-gram matching
- **GTM, METEOR**: Based on the F-measure
- **OI**: Lexical Overlap measure
- **ULC**: *U*niform *L*inear *C*ombination of the previous

Evaluation

Framework II

Heterogeneous set of automatic metrics

– **Syntactic metrics**

- **SP-O_{p}, SP-O_{c}, SP-pNIST**: Lexical overlap according to the *p*art-of-speech or *c*hunk and the NIST score over these elements (*S*hallow *P*arsing)
- **CP-O_{p}, CP-O_{c}, CP-STM9**: Lexical overlap among *p*art-of-speech or *c*onstituents of constituency parse trees (*C*onstituency *P*arsing)
- **ULC**: *U*niform *L*inear *C*ombination of the previous

Evaluation

ca-es Translation

Catalan to Spanish

	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	OI	ULC
DCU1	15.24	12.49	13.25	76.73	12.09	72.75	83.80	83.37	83.70	77.84
DCU2	15.15	12.41	13.21	76.52	12.09	72.18	83.76	83.70	83.56	77.86
DCU3	**14.59**	**11.74**	**12.50**	**77.70**	**12.16**	**73.37**	**84.63**	**84.45**	**84.64**	**79.67**
UPC1	20.17	16.40	19.42	68.20	11.22	62.71	78.46	77.31	74.72	63.82
UPC2	25.10	17.09	22.25	63.12	10.93	57.92	76.44	75.56	73.76	57.45

	CP-Oc(*)	CP-Op(*)	CP-STM9	SP-Op(*)	SP-Oc(*)	SP-pNIST	ULC
DCU1	80.92	81.40	74.10	81.78	83.03	10.87	99.24
DCU2	80.71	81.50	74.19	81.66	82.80	10.90	99.22
DCU3	**81.64**	**82.27**	**74.48**	**82.40**	**83.73**	**10.92**	**100.00**
UPC1	68.52	70.93	58.74	70.95	71.97	9.36	84.47
UPC2	70.59	72.89	63.05	73.01	73.99	10.01	88.06

Evaluation

ca-es Translation

SMT+Apertium

Catalan to Spanish

	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	OI	ULC
DCU1	15.24	12.49	13.25	76.73	12.09	72.75	83.80	83.37	83.70	77.84
DCU2	15.15	12.41	13.21	76.52	12.09	72.18	83.76	83.70	83.56	77.86
DCU3	**14.59**	**11.74**	**12.50**	**77.70**	**12.16**	**73.37**	**84.63**	**84.45**	**84.64**	**79.67**
UPC1	20.17	16.40	19.42	68.20	11.22	62.71	78.46	77.31	74.72	63.82
UPC2	25.10	17.09	22.25	63.12	10.93	57.92	76.44	75.56	73.76	57.45

	CP-Oc(*)	CP-Op(*)	CP-STM9	SP-Op(*)	SP-Oc(*)	SP-pNIST	ULC
DCU1	80.92	81.40	74.10	81.78	83.03	10.87	99.24
DCU2	80.71	81.50	74.19	81.66	82.80	10.90	99.22
DCU3	**81.64**	**82.27**	**74.48**	**82.40**	**83.73**	**10.92**	**100.00**
UPC1	68.52	70.93	58.74	70.95	71.97	9.36	84.47
UPC2	70.59	72.89	63.05	73.01	73.99	10.01	88.06

Evaluation

ca-es Translation

Spanish to Catalan

	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	OI	ULC
DCU1	16.70	12.42	14.46	75.79	11.88	70.45	52.08	82.65	82.88	66.23
DCU2	15.17	11.71	**13.21**	77.75	11.96	72.65	53.44	**83.32**	**83.96**	69.98
DCU3	17.09	13.08	14.70	75.25	11.85	70.34	51.73	82.26	82.46	64.94
UPC1	**14.35**	**11.25**	13.63	**77.93**	**12.04**	**72.69**	**53.98**	82.19	83.18	**70.56**

	CP-Oc(*)	CP-Op(*)	CP-STM9	SP-Op(*)	SP-Oc(*)	SP-pNIST	ULC
DCU1	–	–	–	80.77	82.10	10.78	98.41
DCU2	–	–	–	**82.13**	**83.14**	10.88	**99.67**
DCU3	–	–	–	80.19	81.42	10.75	97.81
UPC1	–	–	–	81.59	82.02	**10.99**	99.33

Evaluation

ca-es Translation

SMT & SMT!

Spanish to Catalan

	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	OI	ULC
DCU1	16.70	12.42	14.46	75.79	11.88	70.45	52.08	82.65	82.88	66.23
DCU2	15.17	11.71	**13.21**	77.75	11.96	72.65	53.44	**83.32**	**83.96**	69.98
DCU3	17.09	13.08	14.70	75.25	11.85	70.34	51.73	82.26	82.46	64.94
UPC1	**14.35**	**11.25**	13.63	**77.93**	**12.04**	**72.69**	**53.98**	82.19	83.18	**70.56**

	CP-Oc(*)	CP-Op(*)	CP-STM9	SP-Op(*)	SP-Oc(*)	SP-pNIST	ULC
DCU1	–	–	–	80.77	82.10	10.78	98.41
DCU2	–	–	–	**82.13**	**83.14**	10.88	**99.67**
DCU3	–	–	–	80.19	81.42	10.75	97.81
UPC1	–	–	–	81.59	82.02	**10.99**	99.33

Evaluation

eu-es Translation

Basque to Spanish

	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	OI	ULC
DCU1	62.19	44.72	56.37	25.30	6.46	32.70	45.71	34.20	44.48	59.78
DCU2	61.24	44.95	55.35	25.30	6.53	33.14	46.12	34.61	44.92	60.63
DCU3	**61.04**	44.78	54.99	25.44	6.56	33.34	46.32	35.31	45.50	61.31
EHU1	61.53	**38.17**	**52.96**	**28.61**	**6.94**	**34.53**	**50.57**	**40.80**	**51.12**	**69.13**

	CP-Oc(*)	CP-Op(*)	CP-STM9	SP-Op(*)	SP-Oc(*)	SP-pNIST	ULC
DCU1	36.82	38.54	29.67	40.94	43.43	5.24	87.99
DCU2	37.13	38.84	29.77	41.16	43.67	5.23	88.40
DCU3	37.71	39.32	30.11	41.69	44.20	5.27	89.45
EHU1	**43.26**	**45.19**	**33.59**	**47.42**	**49.8**	**5.48**	**100.00**

Evaluation

eu-es Translation

SMT

Basque to Spanish

	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	OI	ULC
DCU1	62.19	44.72	56.37	25.30	6.46	32.70	45.71	34.20	44.48	59.78
DCU2	61.24	44.95	55.35	25.30	6.53	33.14	46.12	34.61	44.92	60.63
DCU3	**61.04**	44.78	54.99	25.44	6.56	33.34	46.32	35.31	45.50	61.31
EHU1	61.53	**38.17**	**52.96**	**28.61**	**6.94**	**34.53**	**50.57**	**40.80**	**51.12**	**69.13**

	CP-Oc(*)	CP-Op(*)	CP-STM9	SP-Op(*)	SP-Oc(*)	SP-pNIST	ULC
DCU1	36.82	38.54	29.67	40.94	43.43	5.24	87.99
DCU2	37.13	38.84	29.77	41.16	43.67	5.23	88.40
DCU3	37.71	39.32	30.11	41.69	44.20	5.27	89.45
EHU1	**43.26**	**45.19**	**33.59**	**47.42**	**49.8**	**5.48**	**100.00**

Evaluation

eu-es Translation

Spanish to Basque

	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	OI	ULC
DCU1	61.48	47.56	55.81	23.22	5.96	32.45	40.00	29.92	42.87	66.27
DCU2	**61.06**	46.27	**55.17**	**24.44**	6.12	**33.17**	41.18	31.95	44.29	69.18
DCU3	61.77	47.30	56.07	23.42	5.96	32.48	40.12	30.38	43.00	66.56
EHU1	62.00	**45.04**	56.06	24.34	**6.14**	**33.17**	**41.98**	**32.22**	**45.07**	**69.63**
EHU2	66.43	50.13	62.46	19.54	5.29	29.29	36.36	23.30	38.15	55.33

Evaluation

eu-es Translation

SMT & Comb.

Spanish to Basque

	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	OI	ULC
DCU1	61.48	47.56	55.81	23.22	5.96	32.45	40.00	29.92	42.87	66.27
DCU2	**61.06**	46.27	**55.17**	**24.44**	6.12	**33.17**	41.18	31.95	44.29	69.18
DCU3	61.77	47.30	56.07	23.42	5.96	32.48	40.12	30.38	43.00	66.56
EHU1	62.00	**45.04**	56.06	24.34	**6.14**	**33.17**	**41.98**	**32.22**	**45.07**	**69.63**
EHU2	66.43	50.13	62.46	19.54	5.29	29.29	36.36	23.30	38.15	55.33

Evaluation

pt-es Translation

Portuguese to Spanish

	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	OI	ULC
DCU1	40.51	33.22	37.39	43.36	8.70	42.69	58.85	52.59	58.77	65.48
DCU2	39.86	33.41	36.87	43.67	8.74	43.28	59.12	52.86	58.96	66.17
DCU3	**39.08**	**33.09**	**36.11**	**44.28**	**8.83**	**43.90**	**59.89**	**53.61**	**59.54**	**67.54**

	CP-Oc(*)	CP-Op(*)	CP-STM9	SP-Op(*)	SP-Oc(*)	SP-pNIST	ULC
DCU1	53.57	55.48	44.99	57.32	59.06	8.15	98.51
DCU2	53.85	55.66	45.30	57.48	59.24	8.17	98.89
DCU3	**54.53**	**56.28**	**45.96**	**58.06**	**59.92**	**8.23**	**100.00**

Evaluation

pt-es Translation

Comb. 3 sys

Portuguese to Spanish

	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	OI	ULC
DCU1	40.51	33.22	37.39	43.36	8.70	42.69	58.85	52.59	58.77	65.48
DCU2	39.86	33.41	36.87	43.67	8.74	43.28	59.12	52.86	58.96	66.17
DCU3	**39.08**	**33.09**	**36.11**	**44.28**	**8.83**	**43.90**	**59.89**	**53.61**	**59.54**	**67.54**

	CP-Oc(*)	CP-Op(*)	CP-STM9	SP-Op(*)	SP-Oc(*)	SP-pNIST	ULC
DCU1	53.57	55.48	44.99	57.32	59.06	8.15	98.51
DCU2	53.85	55.66	45.30	57.48	59.24	8.17	98.89
DCU3	**54.53**	**56.28**	**45.96**	**58.06**	**59.92**	**8.23**	**100.00**

Evaluation

pt-es Translation

Spanish to Portuguese

	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	OI	ULC
DCU1	47.68	39.40	44.45	36.13	7.57	37.71	53.78	44.10	52.38	65.27
DCU2	**46.51**	36.67	**43.08**	**37.25**	**7.77**	**38.30**	**54.15**	**45.24**	**53.57**	**68.05**
DCU3	47.04	**36.51**	43.39	36.94	7.76	38.14	53.71	45.19	53.39	67.61

Evaluation

pt-es Translation

Moses+cdec

Spanish to Portuguese

	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	OI	ULC
DCU1	47.68	39.40	44.45	36.13	7.57	37.71	53.78	44.10	52.38	65.27
DCU2	**46.51**	36.67	**43.08**	**37.25**	**7.77**	**38.30**	**54.15**	**45.24**	**53.57**	**68.05**
DCU3	47.04	**36.51**	43.39	36.94	7.76	38.14	53.71	45.19	53.39	67.61

Conclusions

Overview

Conclusions

Summary

- We have created **corpora** of parallel formal tweets in the *ca-es*, *eu-es*, *gl-es* & *pt-es* language pairs
- The translation task resulted **easy** in terms of the output of the automatic evaluation
- Specialised **SMT systems** seem more appropriate than RBMT systems for the task

Conclusions

Next steps

- Agree on **how** a tweet (either formal or informal) must be translated and evaluated
- Widen the task to include also **informal tweets**
- Include **English** in future editions to attract more participants

Conclusions

Thanks!

Now, more TweetMT

Conclusions

References I

Alegria, I. n., Aranberri, N., España Bonet, C., Gamallo, P., Gonçalo Oliveira, H., Martínez, E., San Vicente, I. n., Toral, A., and Zubiaga, A. (2015).
Overview of tweetmt: A shared task on machine translation of tweets at sepln 2015.
In *TweetMT@SEPLN*, pages 1–12.

Alegria, I. n., Aranberri, N., Fresno, V., Gamallo, P., Padró, L., San Vicente, I. n., Turmo, J., and Zubiaga, A. (2013).
Introducción a la tarea compartida tweet-norm 2013: Normalización léxica de tuits en español.
In *Tweet-Norm@SEPLN*, pages 1–9.

Zubiaga, A., San Vicente, I. n., Gamallo, P., Pichel, J. R., Alegria, I. n., Aranberri, N., Ezeiza, A., and Fresno, V. (2014).
Overview of tweetlid: Tweet language identification at sepln 2014.
TweetLID@SEPLN.