Moving the Needle in NLP Technology for the Processing of Code-Switching Language

Thamar Solorio

Disclaimer

- ★ My training is in computer science
- ★ I do not assume any specific theory of code-switching (CSW)
- ★ Everything that mixes languages is CSW (to me)

My Goals Today

- ★ Increase awareness regarding diversity in linguistic abilities
- ★ Motivate more research in NLP for code-switching data
- ★ Showcase some of the work in my group

Outline of the Talk

- ★ Background in code-switching (CSW)
  - Facts
  - Linguistic levels
  - Typology
- ★ Motivating research in CSW
- ★ Our efforts into NLP for CSW
  - Creating linguistic resources
  - Approaches for CSW data
- ★ Work by others
- ★ Final thoughts

Code-Switching (CSW) Definition

The phenomenon by which multilingual speakers switch back and forth between their common languages.

Involves two or more languages or varieties of a language.

Relevant Code-Switching Factors

- ★ CSW is not random, not deterministic either
- ★ Context (Muysken, 2013)
- ★ Language Proficiency
- ★ Pragmatics

Prof. Barbara E. Bullock and Almeida Jacqueline Toribio

Background: Linguistic Levels of Code-Switching

- ★ Phonological

Background: Linguistic Levels of Code-Switching

- ★ Phonological
- ★ Morphological

Estoy kiteando
(*I am joking*)

Cuando mi novio tweetea
pero no contesta 😌
(*When my boyfriend tweets but doesn't answer*)

Background: Linguistic Levels of Code-Switching

- ★ Phonological
- ★ Morphological
- ★ Lexical

I'm going on semana santa ya teniendo mi income tax porque orita ta muy triste la situacion lmaoo

(*I'm going on Holy Week once I have my income tax because right now the situation is pretty bad lmaoo*)

Background: Linguistic Levels of Code-Switching

- ★ Phonological
- ★ Morphological
- ★ Lexical
- ★ Syntactic

Un poquito después del sage llega otro olor dulzón, casi empalagozo: Chilean night-blooming jasmine

A little bit after the sage, another kind of sweet smell arrives, almost cloying: Chilean night-blooming jasmine

Killer Crónicas

Background: Linguistic Levels of Code-Switching

- ★ Phonological
- ★ Morphological
- ★ Lexical
- ★ Syntactic
- ★ Semantic

Agarrar becas (*get scholarships*)
Agarrar trabajo (*get work*)
Agarrar experiencia (*get experience*)
Agarrar buenas notas (*get good grades*)
Agarrar credito (*get credit*)
Agarrar my Bachelor's (*get my bachelor's*)

Bullock, B.E., J. Serigos, A. J. Toribio, 2020

Background: Linguistic Levels of Code-Switching

- ★ Phonological
- ★ Morphological
- ★ Lexical
- ★ Syntactic
- ★ Semantic
- ★ Discourse/pragmatic

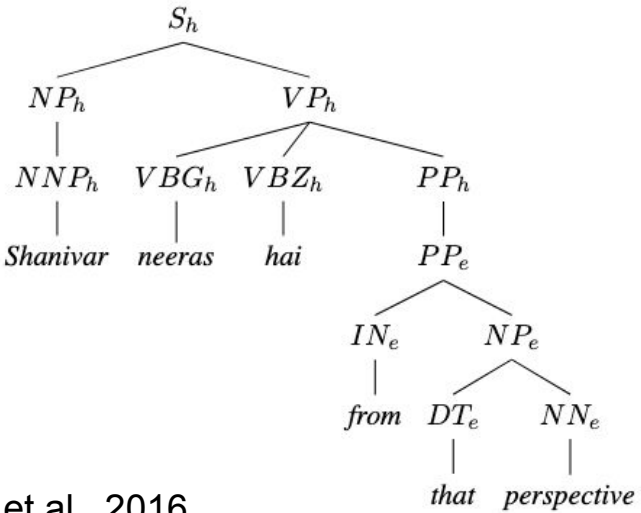
L1 speakers of Hindi tend to switch to Hindi for expressing negative emotions (swearing) (cf. Rudra et al., 2016)

Typology of Code-Switching

Background: Typology of Code-Switching

Muysken (2000)

★ Insertion → Matrix Language (Myers-Scotton, 1993)



Bhat et al., 2016

Background: Typology of Code-Switching

Muysken (2000)

★ Alternation → equivalence principles (MacSwan, 2000; Seba, 1998; Poplack, 1980)

Mi sueño hecho reality: I was going to live in the mero corazón de Cortázar- and Borges-landia.

(*My dream come true: I was going to live right in the heart of Cortázar- and Borges-land.*)

Killer Crónicas

Background: Typology of Code-Switching

Muysken (2000)

★ Congruent Lexicalization

You've got no idea how vinnig I've been slaan-ing this bymekaar together

(*You have no idea how quickly I've been throwing this together.*)

Background: Typology of Code-Switching

★ Ethnolectal (back-flagging)

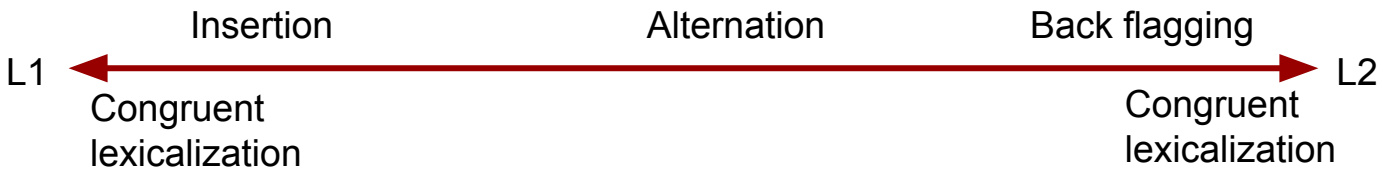
<Ça va>. Why don't you rewire this place and get some regular light switches?

(*It's ok...*)

Guzman et al., 2016

Background: Typology of Code-Switching

Muysken (2000)



Outline of the Talk

- ★ Background in code-switching (CSW)
  - Facts
  - Linguistic levels
  - Typology
- ★ Motivating research in CSW
- ★ Our efforts into NLP for CSW
  - Creating linguistic resources
  - Approaches for CSW data
- ★ Work by others
- ★ Final thoughts

Why is Processing Code-Switching Hard?

Why is Processing Code-Switching Hard?

- ★ It’s not just a mix of languages

Why is Processing Code-Switching Hard?

- ★ It’s not just a mix of languages
- ★ We don’t have linguistic resources

Why is Processing Code-Switching Hard?

- ★ It’s not just a mix of languages
- ★ We don’t have linguistic resources
- ★ Transliteration issues

Hindi-English Tweet:
Keep calm and keep kaam se kaam!! #office #tgif #nametag #buddha #SouvenirFromManali #keepcalm

(keep calm and mind your own business!!)

Nepali-English Tweet:
Youtube ma live re chalcha ki vanni aash garam! Optimistic.

(It’s live on Youtube! Let’s hope it works! Optimistic)

Why is Processing Code-Switching Hard?

- ★ It’s not just a mix of languages
- ★ We don’t have linguistic resources
- ★ Transliteration issues
- ★ It’s spontaneous data (repairs, typos, fluid grammar, etc.)

Why Should we do NLP for CSW Data?

- ★ We want to understand what people say
- ★ Our NLP technologies should accommodate the linguistic abilities of diverse speakers
- ★ Multilingual users prefer bots that code-switch (Bawa et al., 2020)
- ★ Code-switching is a frequent linguistic phenomenon

Outline of the Talk

- ★ Background in code-switching (CSW)
  - Facts
  - Linguistic levels
  - Typology
- ★ Motivating research in CSW
- ★ Our efforts into NLP for CSW
  - Creating linguistic resources
  - Approaches for CSW data
- ★ Work by others
- ★ Final thoughts

Efforts in NLP for CSW

Creating Linguistic Resources

Creating Linguistic Resources

- ★ Language identification
  - Spanish-English
  - Nepali-English
  - Modern Standard Arabic-Egyptian Arabic
- ★ Named entity recognition
  - Spanish-English
  - Modern Standard Arabic Egyptian Arabic

- ★ Sponsors:

Collaborators:
Gustavo Aguilar, Fahad AlGhamdi, Alan Black, Steven Bethard, Elizabeth Blair, Shuguang Chen, Alison Chang, Tanmoy Chakraborty, Amitava Das, Mona Diab, Pascale Fung, Mahmoud Ghoneim, Abdelati Hawwari, Julia Hirschberg, Sudipta Kar, Suraj Maharjan, Giovanni Molina, Suraj Pandey, Parth Patwa, Srinivas P Y K L, Nicolas Rey-Villamizar, Sunayana Sitaram, Victor Soto

Creating Linguistic Resources

- ★ Sentiment analysis
  - Spanish-English
  - Hindi-English
- ★ Machine translation
  - English-Hinglish
  - English-Modern Standard Arabic-Egyptian Arabic
  - English-Spanglish
  - Modern Standard Arabic-Egyptian Arabic-English
  - Spanglish-English

★ Sponsors:

Collaborators:
Gustavo Aguilar, Fahad AlGhamdi, Alan Black, Steven Bethard, Elizabeth Blair, Shuguang Chen, Alison Chang, Tanmoy Chakraborty, Amitava Das, Mona Diab, Pascale Fung, Mahmoud Ghoneim, Abdelati Hawwari, Julia Hirschberg, Sudipta Kar, Suraj Maharjan, Giovanni Molina, Suraj Pandey, Parth Patwa, Srinivas P Y K L, Nicolas Rey-Villamizar, Sunayana Sitaram, Victor Soto

Evaluation Framework for Code-Switching Research

https://ritual.uh.edu/lince/

LinCE: Linguistic Code-Switching Evaluation Framework

Language Pairs	LID	POS	NER	SA	MT
Spanish-English	✓	✓	✓	✓	
Hindi-English	✓	✓	✓		
Nepali-English	✓				
Modern Standard Arabic-Egyptian Arabic	✓		✓		
English-Hinglish					✓
Spanglish-English					✓
English-Spanglish					✓
(Modern Standard Arabic-Egyptian Arabic)-English					✓
English-(Modern Standard Arabic-Egyptian Arabic)					✓

Gustavo Aguilar, Sudipta Kar and Thamar Solorio (2020). LinCE: A Centralized Linguistic Code-Switching Evaluation Benchmark. Proceedings of the Twelfth International Conference on Language Resources and Evaluation, LREC-2020

Advances in NLP for CSW Data

Recent Work

From Monolingual Pretrained Knowledge to CSW Models

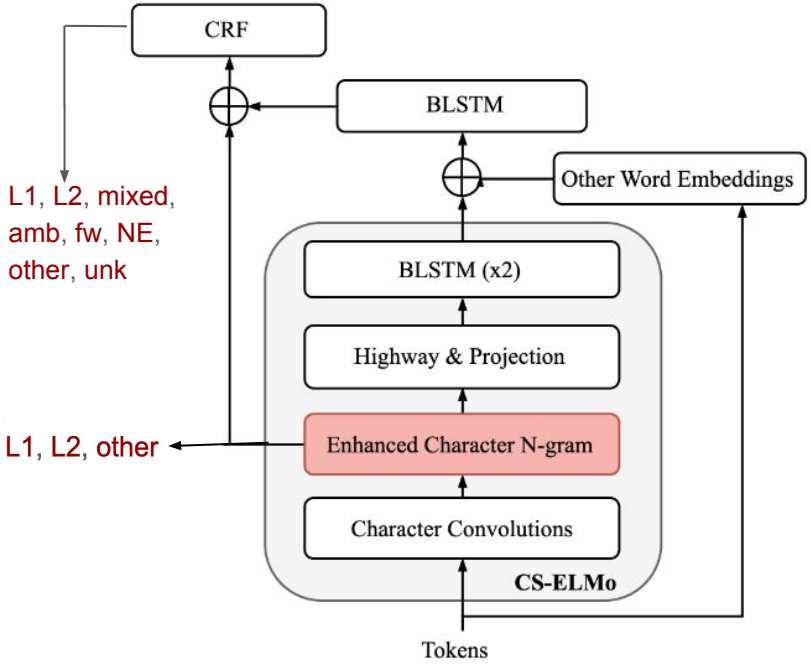
Contextualized embeddings

Transfer Learning From English to CSW Data

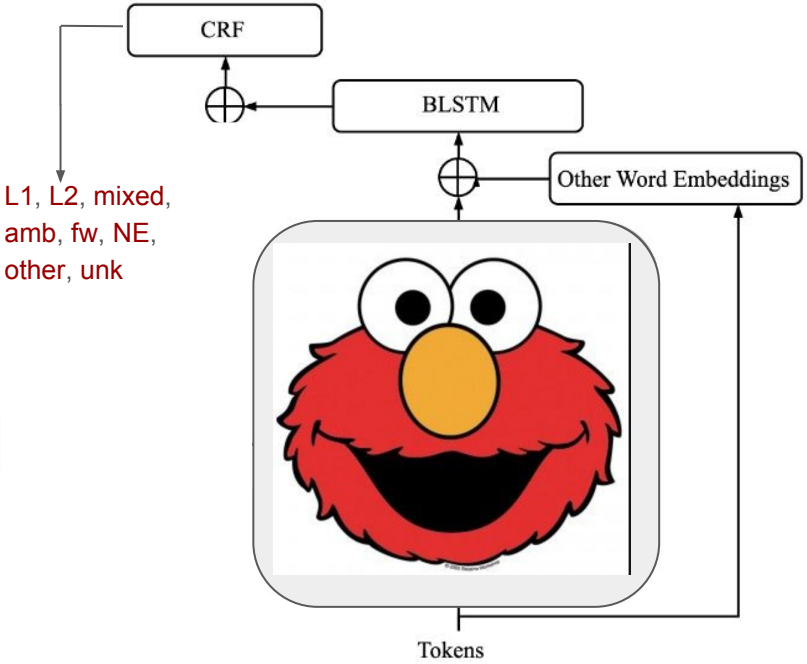
Goals:

- ★ Leverage English pretrained knowledge
- ★ Leverage character-level modelling (morphology) in ELMo (Peters et al., 2018)
- ★ Train a language ID model for code-switching data
  - Labels: Language1, Language2, ambiguous, foreign word, named entity, unknown, other
- ★ Adapt the model for other tasks

Transfer Learning with Strong Morphological Clues

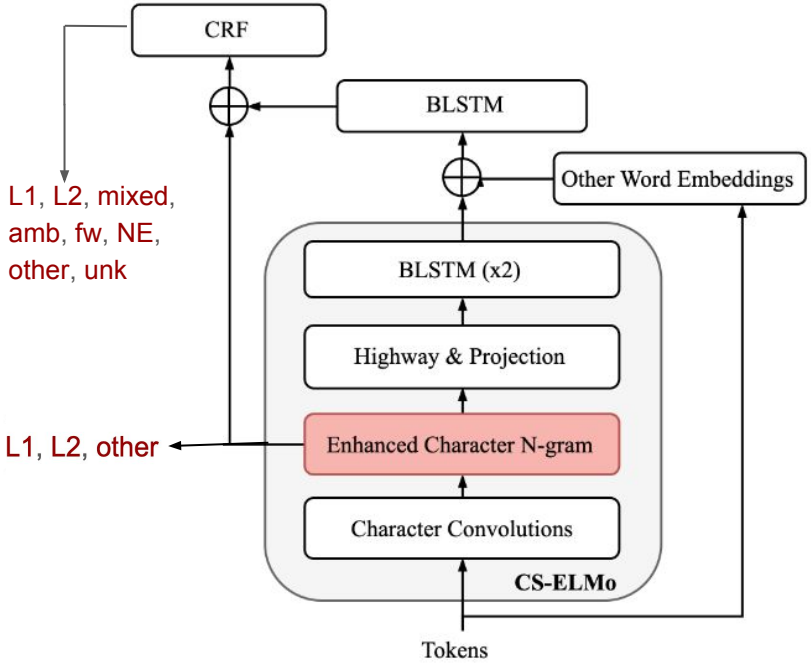


Transfer Learning with Strong Morphological Clues



★ We train CS-ELMo to perform LID: L1, L2, mixed, amb, fw, NE, other, unk

Transfer Learning with Strong Morphological Clues



Enhanced Character N-gram:

- ★ Position embeddings
- ★ Hierarchical attention
- ★ Secondary learning objective:
  - ○ L1, L2, other

$$\mathcal{L}_{task_t} = -\frac{1}{N}\sum_{i}^{N} y_i \, log \, p(y_i|\Theta)$$

$$\mathcal{L} = \mathcal{L}_{task_1} + \beta\mathcal{L}_{task_2} + \lambda \sum_{k}^{|\Theta|} w_k^2$$

Transfer Learning with Strong Morphological Clues

Results

NER System (Spanglish)	Dev F_1	Test F_1
mBERT	61.11	64.56

Dataset from Aguilar et al. (2018)

POS Tagging Model (Hinglish)	Dev F_1	Test F_1
mBERT	86.84	84.70

Dataset from Singh et al. (2018)

Gustavo Aguilar & Thamar Solorio (2020). From English to Code-Switching: Transfer Learning with Strong Morphological Clues. In Proceedings of The 58th annual meeting of the Association for Computational Linguistics, ACL.

Transfer Learning with Strong Morphological Clues

Results

NER System (Spanglish)	Dev F_1	Test F_1
mBERT	61.11	64.56
ELMo+ BLSTM + CRF	59.91	63.53

Dataset from Aguilar et al. (2018)

POS Tagging Model (Hinglish)	Dev F_1	Test F_1
mBERT	86.84	84.70
ELMo+ BLSTM + CRF	87.42	88.12

Dataset from Singh et al. (2018)

Gustavo Aguilar & Thamar Solorio (2020). From English to Code-Switching: Transfer Learning with Strong Morphological Clues. In Proceedings of The 58th annual meeting of the Association for Computational Linguistics, ACL.

Transfer Learning with Strong Morphological Clues

Results

NER System (Spanglish)	Dev F_1	Test F_1
mBERT	61.11	64.56
ELMo+ BLSTM + CRF	59.91	63.53
Prev. SOTA (Winata et al., 2019)		66.63

Dataset from Aguilar et al. (2018)

POS Tagging Model (Hinglish)	Dev F_1	Test F_1
mBERT	86.84	84.70
ELMo+ BLSTM + CRF	87.42	88.12
Prev. SOTA (Singh et al., 2018)		90.20

Dataset from Singh et al. (2018)

Gustavo Aguilar & Thamar Solorio (2020). From English to Code-Switching: Transfer Learning with Strong Morphological Clues. In Proceedings of The 58th annual meeting of the Association for Computational Linguistics, ACL.

Transfer Learning with Strong Morphological Clues

Results

NER System (Spanglish)	Dev F_1	Test F_1
mBERT	61.11	64.56
ELMo+ BLSTM + CRF	59.91	63.53
Prev. SOTA (Winata et al., 2019)		66.63
CS-ELMo	64.39	**67.96**

Dataset from Aguilar et al. (2018)

POS Tagging Model (Hinglish)	Dev F_1	Test F_1
mBERT	86.84	84.70
ELMo+ BLSTM + CRF	87.42	88.12
Prev. SOTA (Singh et al., 2018)		90.20
CS-ELMo	90.37	**91.03**

Dataset from Singh et al. (2018)

Gustavo Aguilar & Thamar Solorio (2020). From English to Code-Switching: Transfer Learning with Strong Morphological Clues. In Proceedings of The 58th annual meeting of the Association for Computational Linguistics, ACL.

From Multilingual Models to Code-Switching

Extending pre-trained multilingual models through improved character n-gram representation learning

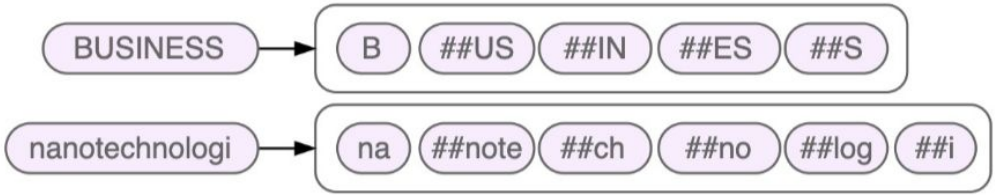
Multilingual Transformer Models

- ★ Trained on >100s of languages
- ★ Are they the solution to Code-Switching data?
- ★ Performance of mBERT drops ~ 40% with transliterated data (Pires et al., 2019)
- ★ Multilingual models (mBERT) < trained on CSW < trained on real CSW data (Santy et al., 2021)
- ★ Smaller CSW specific models are competitive to XLM-R (Winata et al., 2021)
- ★ Our research question: Would something similar to CS-ELMo be feasible?

Tokenization: Byte Pair Encoding

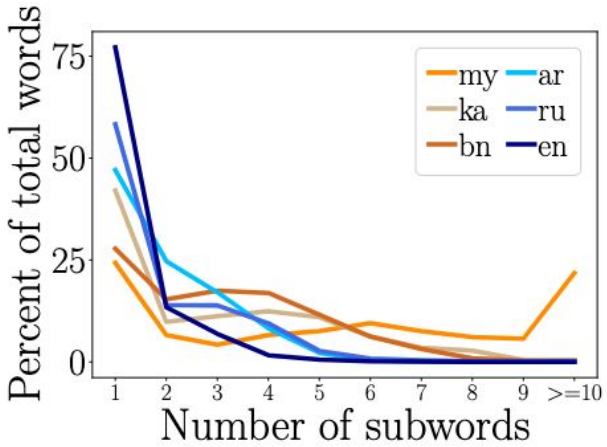
Tokenization: Byte Pair Encoding

- Models like BERT, and XLM-RoBERTa rely on the byte pair encoding (Senrich et al., 2016) or similar:
  - Given a subword s_i, BPE returns a fixed embedding e_i



Undesirable Features in BPE

- ★ Disregards morphological knowledge
- ★ Lack robustness
- ★ Uneven performance in different languages

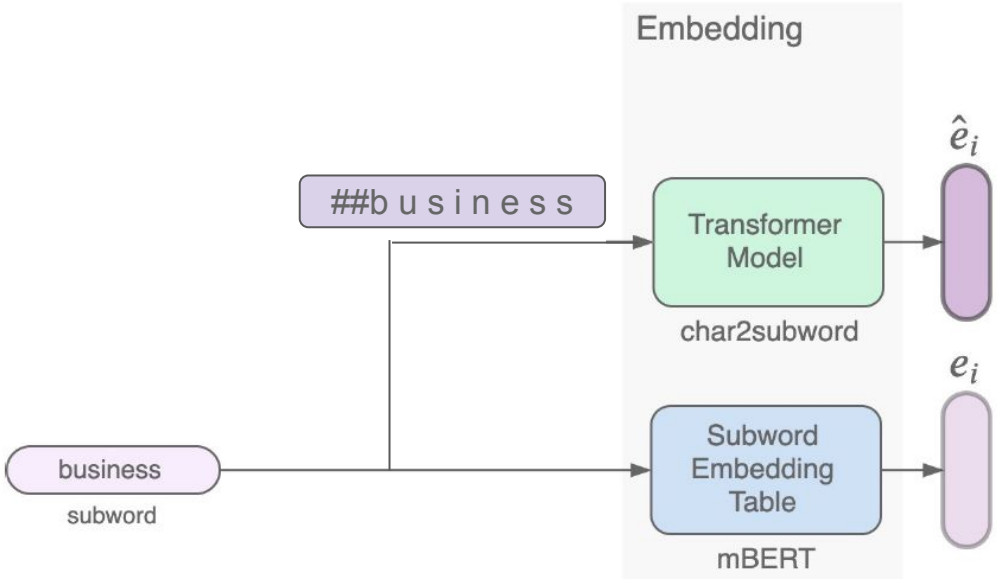


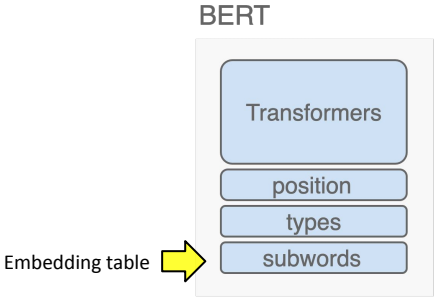
Wang et al., 2021

Increasing Tokenization Robustness

- ★ Our goal is to alleviate some of the shortcomings in BPE
- ★ We train f_Θ to approximate e_i
- ★ We add noise to the input

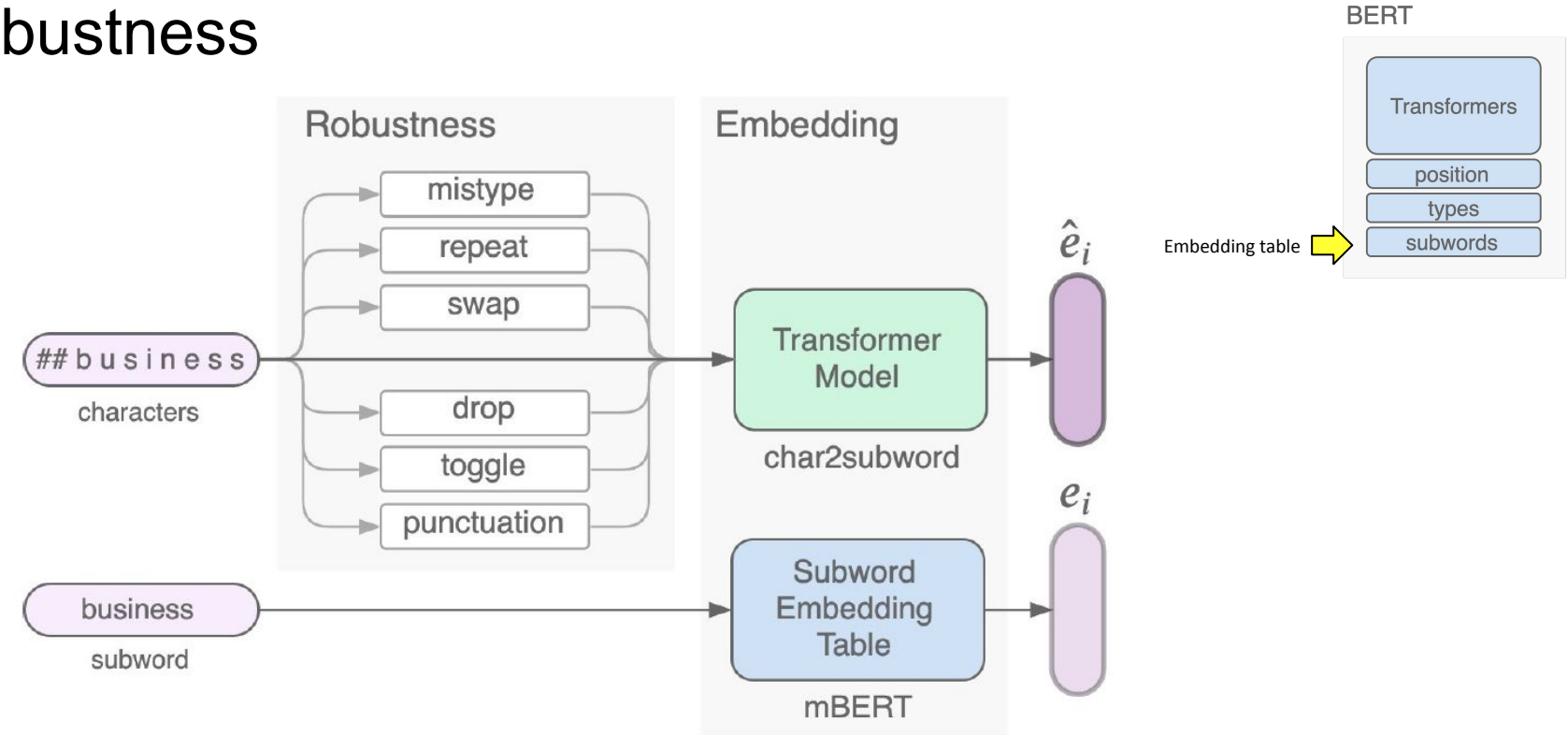
Simulating the Subword Embedding Table





$$\mathcal{L}(\boldsymbol{c}_i, \boldsymbol{e}_i, \boldsymbol{y}_i, f_\theta) = \mathcal{L}_{cos}(\boldsymbol{e}_i, f_\theta(c_i)) + L^2(\boldsymbol{e}_i, f_\theta(c_i)) + \mathcal{L}_{nbr}(\boldsymbol{e}_i, f_\theta(c_i)) + \mathcal{L}_{ce}(\boldsymbol{y}_i, f_\theta(c_i))$$

Simulating the Subword Embedding Table: Adding Robustness



Most similar words

business

1 - business (1.0000)
2 - Business (0.6129)
3 - бизнес (0.5244)
4 - businesses (0.4772)
5 - bisnis (0.4600)
6 - industry (0.4460)
7 - negocios (0.4288)
8 - negocio (0.4169)
9 - enterprise (0.4144)
10 - corporate (0.4065)

mBERT

business

1 - business (0.8209)
2 - Business (0.5037)
3 - businesses (0.4350)
4 - бизнес (0.4038)
5 - bisnis (0.3800)
6 - service (0.3659)
7 - corporate (0.3630)
8 - enterprise (0.3583)
9 - industry (0.3557)
10 - negocios (0.3557)

char2subword

business

1 - business (0.8047)
2 - Business (0.6161)
3 - businesses (0.5348)
4 - бизнес (0.4361)
5 - negocios (0.3928)
6 - enterprise (0.3882)
7 - industry (0.3878)
8 - companies (0.3858)
9 - bisnis (0.3851)
10 - corporate (0.3849)

char2subword + noise

b u s s i n n e s
1 - business (0.5189)
2 - businesses (0.3924)
3 - businessman (0.3458)
4 - бизнес (0.3369)
5 - Business (0.3312)

b i s i n e s s
1 - bisnis (0.5208)
2 - business (0.4482)
3 - businesses (0.4329)
4 - (0.4049) الشركات
5 - бизнес (0.4039)

b u s I N e s s
1 - Ichthyology (0.2644)
2 - censusindia (0.2550)
3 - Rechtsanwalt (0.2369)
4 - Melastomataceae (0.2332)
5 - SQL (0.2328)

b U s i n e s s
1 - Machines (0.3943)
2 - Solutions (0.3729)
3 - Reviews (0.3642-)
4 - Products (0.3517)
5 - Recordings (0.3465)

B U S I N E S S
1 - Historische (0.2957)
2 - SDSS (0.2753)
3 - IAU (0.2747)
4 - UNESCO (0.2737)
5 - ESA (0.2718)

char2subword

b u s s i n n e s
1 - business (0.6679)
2 - businesses (0.6159)
3 - Business (0.4873)
4 - companies (0.4120)
5 - negocios (0.3880)

b i s i n e s s
1 - business (0.5945)
2 - businesses (0.5408)
3 - bisnis (0.4915)
4 - negocios (0.4884)
5 - customers (0.4297)

b u s I N e s s
1 - business (0.7967)
2 - Business (0.6148)
3 - businesses (0.5304)
4 - бизнес (0.4312)
5 - companies (0.3879)

b U s i n e s s
1 - business (0.8044)
2 - Business (0.6076)
3 - businesses (0.5352)
4 - бизнес (0.4289)
5 - corporate (0.3888)

B U S I N E S S
1 - Business (0.5313)
2 - business (0.3264)
3 - Marketing (0.3180)
4 - Corporate (0.3126)
5 - Communications (0.3049)

char2subword + noise

Results on the LinCE Benchmark

Method	Adaptation	Avg	LID (W. F_1)				POS (Acc.)		NER (F_1)			SA (W Acc.)
			es-en	hi-en	ne-en	msa-arz	es-en	hi-en	es-en	hi-en	msa-arz	es-en
ELMo	N/A	79.52	97.93	95.43	95.90	86.53	96.34	86.71	52.58	68.79	56.68	58.32
mBERT	N/A	82.23	**98.36**	94.24	**96.32**	**91.55**	**97.07**	86.30	64.05	72.57	65.39	56.43
Hybrid	Pre-trained*	**83.03**	98.33	**96.23**	96.19	91.19	96.88	**88.23**	**64.65**	**73.38**	**66.13**	**59.07**

* Statistically significant with respect to the mBERT baseline, with p-value < 0.01 in student's t-test (Dror et al., 2018).

G Aguilar, B McCann, T Niu, N Rajani, N Keskar, T Solorio (2020). Char2Subword: Extending the Subword Embedding Space from Pre-trained Models Using Robust Character Compositionality, arXiv preprint arXiv:2010.12730

Outline of the Talk

- ★ Background in code-switching (CSW)
  - ○ Facts
  - ○ Linguistic levels
  - ○ Typology
- ★ Motivating research in CSW
- ★ Our efforts into NLP for CSW
  - ○ Creating linguistic resources
  - ○ Approaches for CSW data
- ★ Work by others
- ★ Final thoughts

Work on CSW by Others

- ★ MSR India: Code-switching modelling & generation (Pratapa et al., 2018; Zaki et al., 2021), text to speech (Sitaram & Black, 2016), GLUECoS (Khanuja et al., 2020)
- ★ Özlem Çetinoglu: Challenges in processing CSW data (Çetinoğlu et al., 2016) Lang Id (Mager et al., 2019), Morphological Tagging (Özatęs and Çetinoğlu, 2021)
- ★ Amitava Das: Sentiment analysis (Patra et al., 2018; Patwa et al., 2020) (Hindi-English), complexity of CSW data (Gambäck & Das, 2016)
- ★ Mona Diab: Language identification (Elfardy et al., 2012,2014), resources for CSW with Arabic languages.
- ★ Pascale Fung: Hierarchical meta embeddings (Winata et al., 2019a,b)
- ★ Others too….

Outline of the Talk

- ★ Background in code-switching (CSW)
  - ○ Facts
  - ○ Linguistic levels
  - ○ Typology
- ★ Motivating research in CSW
- ★ Our efforts into NLP for CSW
  - ○ Creating linguistic resources
  - ○ Approaches for CSW data
- ★ Work by others
- ★ Final thoughts

What is Next?

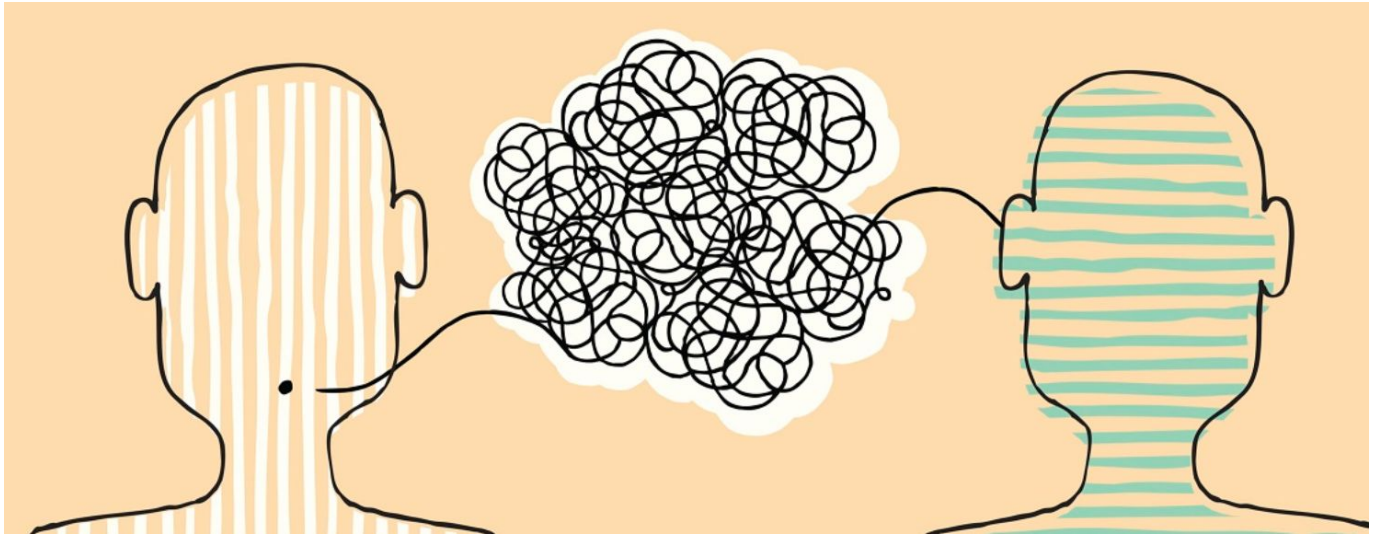
- ★ More CSW resources:
  - Deeper syntactic annotations (dependency parsing)
  - Different language combinations
  - 2+ languages and eventually *n* language combinations
- ★ More research for non-English data
- ★ Reach out to linguists, socio-linguists, language acquisition, language development experts!

Special Thanks to RiTUAL Members!

Interested in Code-Switching?

Fifth Workshop on Computational Approaches to Linguistic Code-Switching (CALCS 2021)

When: June 11th from 8:00-18:15 (GMT-5)



Gracias por su atención!

Visit our website:

ritual.uh.edu

LinCE website: ritual.uh.edu/lince

Twitter: @thamar_solorio

Email: thamar.solorio@gmail.com