Impact of Spanish Dialect in Deep Learning Next Sentence Predictors

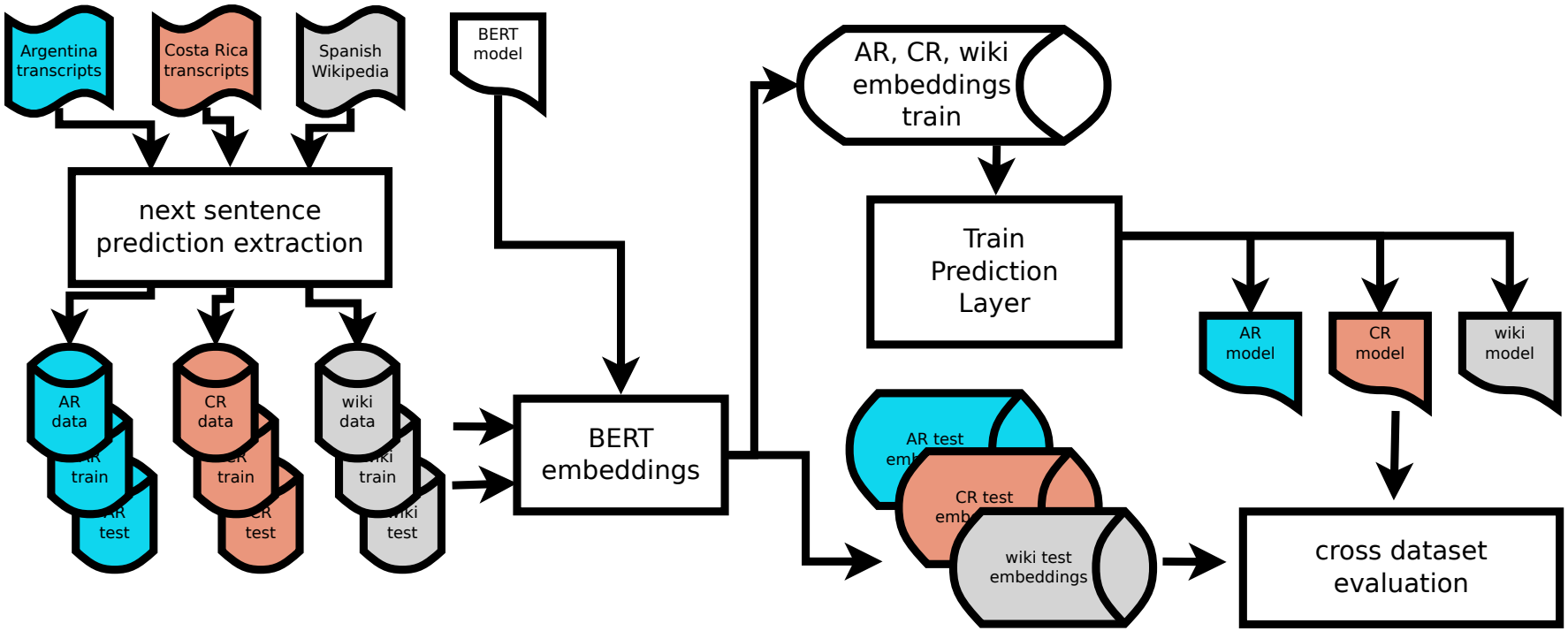
Pablo Duboue

Textualization Software Ltd.
Vancouver, CANADA

Conferencia Latinoamericana de Informatica, 2019

Outline.

One Slide Summary.



https://github.com/IE4OpenData/clei2019dialect

Outline

About the speaker.

- Computer Science, Cordoba, Argentina
  - Years 1994-1998
  - Thesis in Spanish parsing using Haskell
- PhD in Computer Science, Columbia University, NY
  - Years 1999-2005
  - Natural Language Generation (NLP / AI)
  - Thesis on biography generation
- IBM Research Watson: Deep QA - Watson - Jeopardy! Show
- Consulting company in Vancouver (in Montreal from 2010-2016)
- TW @pablodubouе / DrDub (GitHub) / Google Scholar: https://scholar.google.ca/citations?user=Exngg_MAAAAJ (1,500 citations)

Impact of Mandatory Representation.

- Long term research question:
  - Do mandatory gender quotas impact parliamentary behaviour?
  - Are the women in parliament listened and their ideas have similar impact as their male counterparts?
- This work:
  - Analyzing whether solutions based on data from one country might work when moving to a different country

Voz y Voto.

- Voz y Voto project is a study of group participation in the proceedings of government assemblies, both in terms of speaking and being addressed by other speakers.
- Current demo was put together at a hackathon called Hack(at)ONG in Cordoba, Argentina in September, 2016.
  - A topic of interest for the FUNDEPS NGO: `http://www.fundeps.org/en/home/`
- System is a combination of shell scripts, pandoc, iconv, perl and OpenNLP.
  - The gender identification in names is based on the Spanish Wikipedia list of unambiguous male and female names.
  - The entity linking (person disambiguation) component uses a heuristic point system
  - Topic identification per session uses TF*IDF.
- A live web-graph using plot.ly can be seen at `http://vozyvoto.ie4opendata.org/demo.html`.

Impact of Dialect.

- Dialect-ID is much harder than lang-id
  [Ciobanu and Dinu, 2016, Radford and Gallé, 2016]
- English: less confined to regional but rather social groups
  - [Tatman, 2017] robust differences in accuracy for speakers from Scotland and US Deep South
  - [Blodgett and O'Connor, 2017] tweets by African American less likely to be identified as English

Impact of Dialect.

- Arabic: amalgamation of 20+ dialects [Diab and Habash, 2014] (Moroccan, Egyptian or Saudi Arabian)
  - Standard Arabic dictated by the media, and most NLP tools cater to that dialect [Shaalan et al., 2018]
  - Dialects processing is challenging [Shoufan and Alameri, 2015]
  - Dialect-ID is key for spoken interfaces [Lei and Hansen, 2009]
- Chinese: ideograms allows for non-mutually intelligible dialects
  - [Zhang, 2017] studied difference in written form by using corpus linguistics and statistical methods
  - [Peng et al., 2015]: transliterating of foreign words across dialects accounts for a 16% drop in F-measure if not addressed
  - [Xu et al., 2016] Dialect-ID using character n-gram is unsuccessful. SOTA achieves 82% accuracy over 6 dialects.
  - Some dialects are not mutual intelligible and have parallel corpora [Wong et al., 2017, Xu et al., 2018].

Impact of Dialect.

- Spanish: also been the focus of dialect analysis, particularly rural dialects [de Benito Moreno et al., 2016].
  - [Bogantes et al., 2016] analyzed the impact of four dialects (Colombia, Costa Rica, Mexico and Peru) in Multi-Word Expressions (MWE)
  - [Sanchez-Perez et al., 2017] worked on dialect-id and author-id in 8 varieties of the news Spanish (Argentinian, Mexican, Colombian, Chilean, Venezuelan, Panamanian, Guatemalan, and Peninsular Spanish) . Combination of character n-grams and lexical features was best for dialect-id
- Portuguese
  - [Fonseca et al., 2015, Hartmann et al., 2017] dialect specific word embeddings

Impact of sub-language for NLP tasks.

- The impact of sub-language for NLP tasks is well studied for domain specific areas
  - Part-of-Speech tagging [Ferraro et al., 2013] for medical domain
    [Baud et al., 1992, Rzhetsky et al., 2004, Ford et al., 2016] and software engineering [Ferrari et al., 2017].
  - Sub-problems such as negation detection [Miller et al., 2017].
- The domain impact is particularly crucial when dealing with languages with fewer annotated resources [Plank, 2016]. Crucial for Deep Learning [Qu et al., 2015].
- Domain differences can be used to generate "ungrammatical" sentences for parser trained on different domains
  - Analyze the performance for grammatically different sentences [Hashemi and Hwa, 2016].
- Lack of adaptability: a problem for commercial deployment of NLP systems [Dahlmeier, 2017].

Civic applications of NLP technology.

- NLP for political debate
  - Crowd-sourced deliberation [Aitamurto et al., 2016]; feedback collection [Whittle et al., 2010] and digital simulations to engage citizens in political discourse [Poole et al., 2010]
- Parliamentary proceedings are a staple of NLP research [Carpuat, 2014, Wattam et al., 2014, Maekawa et al., 2014].
- Other analysis of political debates: [Onyimadu et al., 2013] sentiment analysis on Canadian hansard.
- Analysis of political discourse key both in political science [Ilie, 2004, Monroe and Schrodt, 2008, Rasiah et al., 2010] and linguistics
  [Bara et al., 2007, Rasiah, 2010, Rodríguez, 2011].

Contextual Embeddings.

- Self-supervised tasks: tasks where labels can be obtained directly from text without the need of annotators.
- Our task: **next sentence prediction**
  - Given two sentences, determine whether they appear contiguous to each other in text, or they are far apart or from different documents
  - Surrogate for **textual coherence**
  - Successfully employed to train sentence orderings for multi-document summarization [Logeswaran et al., 2018]

BERT.

- BERT [Devlin et al., 2018] pre-trained models use next-sentence prediction as their main pre-training objective
  - We profit from BERT performance without heavy fine-tuning.
- BERT is based on CNNs with positional encodings and a deep attention mechanism [Vaswani et al., 2017]
  - Part of the Transformers formalism [Logeswaran and Lee, 2018]
- A training instance consists of
  - a class pseudo-token, [CLS]
  - a sequence of tokens
  - a separator pseudo-token
  - [SEP]
  - a second sequence of tokens
- The tokens are represented using Google's WordPiece tokenizer [Wu et al., 2016]
  - splits words based on frequencies
    - marked with '##' in the examples

BERT Training.

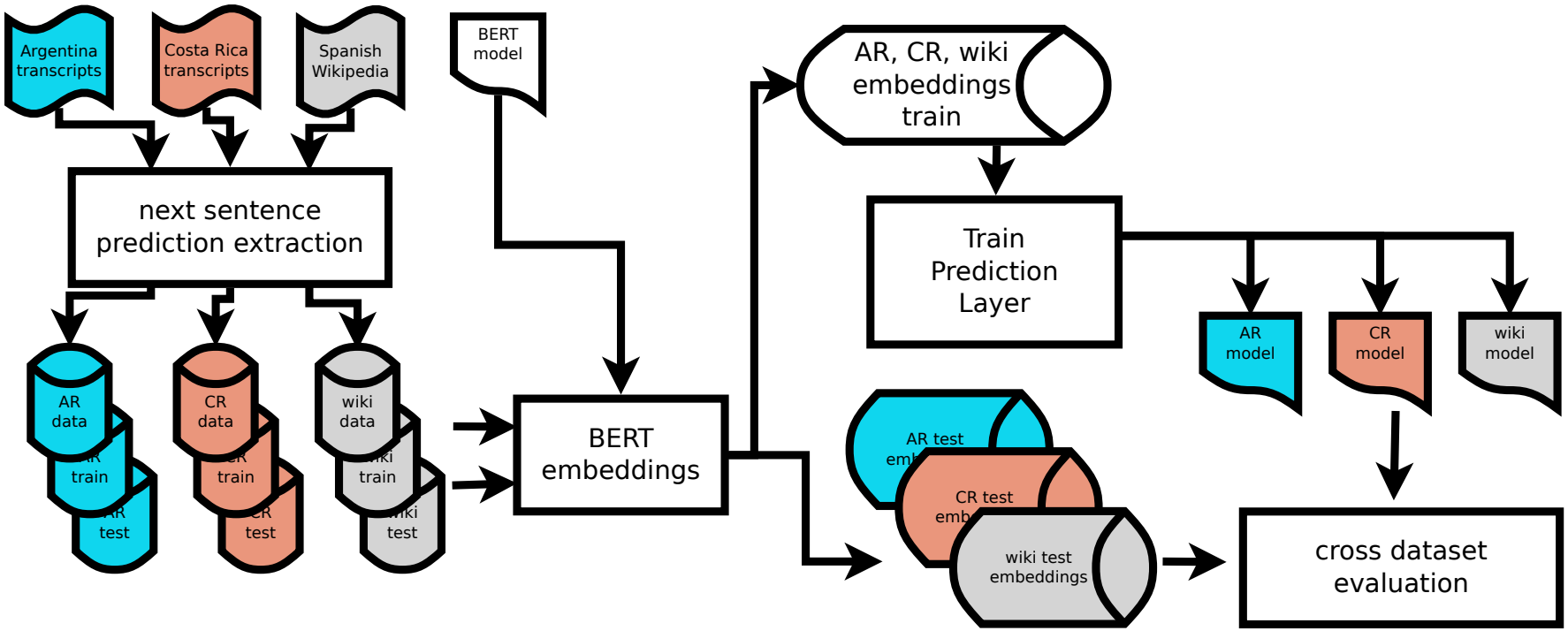
- BERT model is trained to predict with a masked language model
  - to predict missing words explicitly hidden during training
- During training, the BERT embeddings are learned based on the auxiliary task,
  - predicting the isNext value using the embedding of the [CLS] token as input
  - through an extra dense layer of size 768
    - this layer is not distributed with the embeddings, sadly
  - plus predicting the masked words
  - these two predictions is what drives the model to learn the embeddings.
- Therefore, **the embeddings for the [CLS] token are thus trained to predict the isNext class**

BERT Training.

- The BERT model is pre-trained for days on clusters of high-end GPUs with 64Gb of RAM, over billions of words
  - Such models are made available for download by Google LLC
  - Multilingual cased model released in Nov. 2018 (659Mb)
  - Single model handles English, Chinese, Spanish and others
- The two sequences of tokens in BERT input represent two sentences or text segments that can be
  - contiguous (isNext = true)
  - or not (isNext = false)
  - With fine-tuning, the 2 sequences of tokens can be made to be
    - a question and an answer or
    - a piece of text and its paraphrase
    - or ...
  - Fine-tuning currently uses GPUs with 64Gb of RAM
    - As we use the original isNext task, we did not need fine-tuning

Outline

One Slide Summary.



https://github.com/IE4OpenData/clei2019dialect

Intuition.

- **The final system takes two sentences and determines how likely are they to follow each other, a surrogate for a coherence metric.**
- Take a Wikipedia sentence s_1 in page P_1 and take the next sentence s_2 in page P_1.
  - Train to say that the pair $\langle s_1, s_2 \rangle$ is coherent.
- Taking a sentence s_1' in page P_2 and a sentence s_2' in page P_3
  - Train that $\langle s_1', s_2' \rangle$ is not coherent.
- Repeat using three different datasets (Wikipedia, AR and CR)

Extracted Data.

Metric	Argentina	Costa Rica
Raw data (incl. formatting)	161Mb	382Mb
Years	2001-2016	1993-2016
Textual data		
Words	2M	64M
Spoken transcripts		
Segments	17,986	9,965
Sentences	25,133	19,099
Avg. Sent. per Segment	1.40	1.92
Words	1.12M	1.24M
Avg. Words per Sent.	44.95	65.30

Example (Argentina) isNext = true.

[CLS] Yo qui ##ero rei ##vind ##ica ##r a la mayoría de los ju ##ece ##s , que son pro ##bos , que trabaja ##n , que se dedica ##n , aunque como en todo ámbito donde participa el ser humano también existen actitud ##es pat ##ológicas , raya ##nas en las irregular ##idades , y para esto es buen ##o que exist ##an instrumentos para que estas conducta ##s sean **[SEP]** La necesidad de que un poder que parece ai ##sla ##do en relación con la participación popular se am ##pl ##í ##e , se debat ##a y se discu ##ta ya constituye un valor en sí mismo , más allá de los dis ##ens ##os y disc ##re ##pan ##cias que pod ##amos tener . [SEP] (**isNext = true**)

Example (Argentina) isNext = false.

[CLS] Señor presidente : sin án ##imo de quer ##er rei ##tera ##r la reunión de la Comisión de Labor Parlament ##aria , deb ##o decir que en realidad fu ##i mos muchos los presidente ##s de bloque que nos mani ##festa ##mos en esa dirección . **[SEP]** Lam ##enta ##blem ent ##e esta com ##puls ##ión que ten ##emos los argentino ##s de llegar tarde siempre a todo no sólo se rep ##ite esta vez ; además , ahora corre ##mos el riesgo de perder una nueva oportunidad de tomar un tema tan violenta ##mente dolor ##oso . [SEP] (**isNext = false**)

BERT as Feature Extractor.

- **We use BERT as a feature extractor**
  - Similar to ELMO feature extractors [Peters et al., 2018]
- It is possible to obtain a BERT embedding (of 3,072 dimensions) per token
  - This is very onerous in terms of storage space
  - Will require about a terabyte of space for each of the systems used in the current experiments
- Instead, **we use the embedding for the [CLS] token**
  - it captures the isNext task
  - we only need to store 3,072 numbers per pair of sentences

Next Sentence Predictor.

- The CLS embedding contains all the info for isNext prediction
  - a second system needs to be trained on top of them
- Our system:

1. Take pairs of sentences, some contiguous (isNext = true), some not (isNext = false)
   - 50% sampling for each class.
2. **Compute BERT embeddings for each pair (3,072 dimensions)**
3. Train a network to predict isNext
   - We used a dense layer sized 64 using GeLU activation [Hendrycks and Gimpel, 2016]
   - similar to BERT auxiliary task
     - a single neuron as output, using sigmoid activation
   - 30 epochs using the Adam optimizer and a binary cross-entropy loss function
   - Heavy dropout with a keep probability of 25%

Experimental Setting.

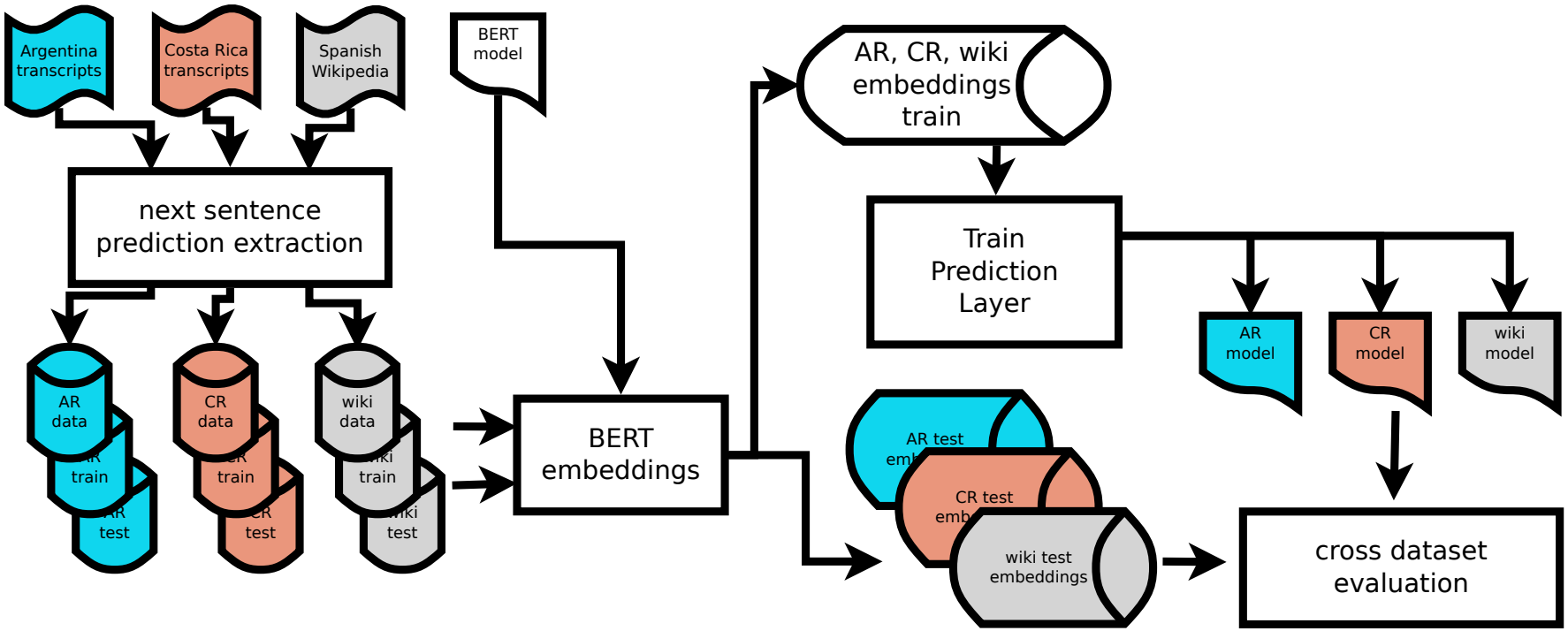
- We split 20% of the sessions for held-out test test:
  - 6.2k sentences for AR
  - 3.6k sentences for CR
  - Sampled 10k sentence pairs for each country
- Over the remaining training data:
  - 18.8k sentences AR
  - 15.5k sentences CR
  - Sampled 100k sentence pairs for each country
- Bacgkound sample of 1M sentences from Spanish Wikipedia
  - 100k sentence pairs
- All pairs have 50% isNext=true and 50% isNext=false

Experimental Setting.

- For each of these five sets
  1. Wikipedia train 100k
  2. AR train 100k
  3. CR train 100k
  4. AR test 10k
  5. CR test 10k
- Extracted the 3,072 BERT embedding using a 6Gb GPU
- From the train datasets:
  - Use the isNext labels+embeddings to train an isNext detector
  - Three trained systems

Outline

One Slide Summary.



https://github.com/IE4OpenData/clei2019dialect

Validation Results.

- Validation accuracies:
  - Wikipedia: 97%
    - Wikipedia is part of the BERT training data
  - AR: 93%
  - CR: 85%.
- Average training pair sizes:
  - Wikipedia: 118 tokens
  - AR: 93 tokens
  - CR: 133 tokens
- **Length difference makes the CR data more challenging**
  - This might be an artifact of the transcription process or a difference in the level of verbosity between the two speaker communities
- Earlier experiments were very unstable
  - We run five different training runs, reporting the mean and the maximum boundary that encompasses all five runs as taken around the mean

Cross Evaluation Results.

Train	Test	Prec	Rec	F1
wiki	AR	52.83 ± 0.28	98.18 ± 0.23	68.70 ± 0.19
wiki	CR	54.14 ± 0.44	98.12 ± 0.31	69.78 ± 0.29
AR	AR	88.46 ± 1.22	93.01 ± 2.48	90.62 ± 0.56
AR	CR	72.16 ± 2.58	90.00 ± 3.26	80.07 ± 0.38
CR	AR	83.74 ± 1.77	82.44 ± 2.76	83.07 ± 0.65
CR	CR	86.39 ± 3.06	81.14 ± 3.38	83.65 ± 0.33

Analysis.

- **Using a general model (Wikipedia) vs. a genre specific model is crucial**, a generic model drops:
  - In Argentina from 83 F-measure to 68
  - In Costa Rica from 80 to 69
  - A relative difference of 17.3% and 12.9%, respectively
- The drop from using a custom model is much higher (Argentina is 21 points of absolute F-measure)
- The numbers have some variability between runs
  - but it is possible to conclude that Costa Rica is closer to standard Spanish as embodied by Wikipedia
    - This result is non-obvious, as the Costa Rica dataset has longer sentences and produced lower validation accuracy
- The validation results track the performance on unseen data
  - 93 to 90 F-measure in Argentina
  - 85 to 83 F-measure in Costa Rica

Cross Evaluation Results.

	Train	
Test	AR	CR
AR	90.6	83.1
CR	80.1	83.7

Analysis.

- When applied to the Argentinian data, the Costa Rican model produced results indistinguishable from the model applied to Costa Rican data
  - It exhibits no dialectical differences
  - **Not a very good model, but equally poor on the two datasets**
- The Argentinian model, however, it is 8% better (relative, 7.55 absolute) on Argentinian data than on Costa Rican data
  - For Argentinian data, it pays off to use a custom model
  - For general use, a model trained on a more neutral dialect is to be preferred
- These are preliminary results that need more dialects for contrast
- This methodology is also very sensitive to changes in the training data and corpus cleaning

Difference Due To Language?

- Are these differences due to differences in language or other differences?
- Looked into a sample of the 306 cases where the AR data had isNext=true and the AR model predicted it correctly **but** the CR model did not
- A clear pattern from that data is the fact that speakers in the Argentinian congress
  - Like to address their words to the president of the chamber
  - Include calls to action in the middle of a speech thus addressed directly to the president's chamber
  - We believe that might be confusing to models where the speakers do not use this style
- These are particularities of a group of speakers that we believe qualify for dialectical differences.

Difference Due To Language?

- More traditional differences include the use of the lexical items such as
  - the verb “desandar” (to go back, to undo), which is not present on the CR data,
  - the word “irrisorio” (risible) that is present twice as much in AR data than in CR data,
  - the word “zapatilla” (sneaker) that is called very differently in other areas of Latinamerica,
  - the word “patronal” (company management, the opposing party in labour union negotiations),
  - “conurbano” (metropolitan area), not present in CR data and
  - “corralito” (small corral) which refers to a banking freeze specific to Argentina.

Difference Due To Language?

(0.0657) En varios medios de comunicación totalmente objetivos, que no distorsionan la información y se expresan sin ningún sentimiento o corrimiento ideológico o propagandístico, como Página/12 o 6, 7, 8, se ha señalado que por culpa de las **patronales** o de los gremios o de la burocracia el empleado rural es el peor pago. [SEP] Es cierto, el empleado rural hoy no percibe grandes salarios, y podría estar gozando de un 37,5 por ciento de aumento de convenio colectivo de trabajo, entre la **patronal** y los empleados, pero no es así porque el Ministerio de Trabajo no homologa más que un 25 por ciento.

Difference Due To Language?

(0.0088) Particularmente, un tema de actualidad -que merece un análisis más exhaustivo, toda vez que la Cámara deberá expresarse sobre su contenido- está vinculado con el decreto 905, referente a la salida de los llamados "**corralito**" y "**corralón**" y al plan que se ha elaborado para ello. [SEP] Por intermedio de las comisiones de Presupuesto y Hacienda y de Finanzas se ha invitado al secretario de Finanzas o al ministro de Economía para que suministren la información que todos deseamos conocer.

Difference Due To Language?

(0.2188) Se transfieren riquezas al exterior despojándonos de la cultura del trabajo, cercenando oportunidades a las nuevas generaciones y también agrandando las asimetrías entre las regiones porque en el puerto quedan las ganancias, fruto de un intercambio desigual con quienes nos venden sus productos manufacturados con trabajo y tecnología incluidos. [SEP] Señor presidente: no es casual que tengamos casi 2 millones de desocupados, 40 por ciento de trabajadores en negro, miles de subsidios y seguros de desempleo, 30 por ciento de trabajadores con ingresos por debajo de sus necesidades e indigentes arracimados en los **conurbanos** a merced del clientelismo.

Term Extraction

- To make these intuitions more systematic, we turned to a simple domain specificity for term extraction [Park et al., 2008] using
  - foreground: the 100k training sentences per country
  - background : the 100k sentences from Wikipedia
- For all nouns, we computed the domain specificity:
  - the ratio of foreground frequencies over the background freqs
  - keep nouns with domain specificity > 3.0 (~3,000 per country)
- We filter them as country or parliamentary specific by
  - looking at the top 1,000 terms for each country and
  - only considering as country specific the ones that appear only in one country list or,
    - if they appear in the other country list, they have a 10-fold difference in score
  - 316 joint, 761 Argentinian and 807 Costa Rican terms

Term Extraction

- The table also includes “poquito”
  - **Costa Ricans are known to be very fond of diminutives**
  - There are 22 terms ending in -ito, -ita marked as Costa Rican-specific terms (compared to 7 for Argentina).
  - It is reassuring to see this phenomenon captured in our data.
- Looking at the 686 Argentinian terms and see how often did they appear on the whole test set:
  - 93.67% of the test sentences contained one of them
- Looking into the 306 pairs that the Argentinian model found but were missed by the Costa Rican model
  - Analyzed whether they had a higher percentage of Argentinian-specific terms: 95.75%
- Analyzing the 3,667 pairs from the Argentina test set that both Argentina and Costa Rican models got correct
  - Incidence of Argentinian specific went down to 89.99%

Term Extraction: Both

spaCy Lemma	Example	Specificity Score
referirme	'referirme'	922.50
quórum	'quórum'	522.50
moción	'Moción', 'mociones', 'moción'	125.50
ponernos	'ponernos'	210.50
votarlo	'votarlo'	186.00
permitirme	'permitirme'	183.00
apelar	'apelo', 'apela', 'apelé'	165.50
inciso	'inciso', 'incisos', 'incisas'	160.23
oficialismo	'oficialismo'	158.44
decirnos	'decirnos'	151.50

Term Extraction: Argentina

spaCy Lemma	Example	Specificity Score
solicitar	'solicito', 'solicitadas', 'solicité'	1294.00
renegociación	'renegociaciones', 'renegociación'	843.00
interbloque	'interbloque'	779.00
modificatorio	'modificatoria', 'modificatorios'	622.00
alícuota	'alícuotas', 'alícuota'	584.00
coparticipación	'coparticipación'	532.00
afirmativo	'afirmativos', 'afirmativa'	449.00
planteos	'planteos'	394.00
apartamiento	'apartamientos', 'apartamiento'	331.25
informarse	'informarse'	285.34

Term Extraction: Costa Rica

spaCy Lemma	Example	Specificity Score
colón	'colones', 'colón'	1574.50
posposición	'posposiciones', 'posposición'	1288.00
talvez	'talvez'	1055.00
costarricense	'costarricense', 'costarricenses'	751.77
señoría	'señoría', 'señorías'	639.17
poquito	'poquita', 'poquitas', 'poquito'	597.00
cafetín	'cafetín'	595.00
irrespeto	'irrespeto'	554.00
solidarismo	'solidarismo'	503.00
portillo	'portillo', 'portillos'	338.00

Remaining Question.

- What if the differences are not due to dialectical differences?
- **Why did the Costa Rican model on Costa Rican data performed below the Argentinian model on Argentinian data?**
  - We know that it has much longer sentences.
  - The impact of this phenomenon can be ablated by splitting the segments into other type of chunks
  - It should be possible to establish by analyzing other regional texts whether they exhibit longer sentence length on average than texts from Argentina
  - It is unclear whether such differences would fall into what is normally considered dialectical differences.
- Another possibility is that the Argentinian data has some particularity that allows the model to easily outperform a general case.
- **Adding more dialects will help**

Summary.

- We have looked into the impact of dialects in the isNext prediction using parliamentary proceedings as training data.
  - **Task-specific differences might be more important than dialect, accounting for up to 21 points of absolute F-measure difference**
  - Dialects also have an impact, accounting for up to 7 points of absolute F-measure difference.
  - Question: which type of dialect is to be preferred for training "generic" models?

`https://github.com/IE4OpenData/clei2019dialect`

Acknowledgements.

- The original Voz y Voto team
  - Annie Ying & Mauricio Korach
- Dr. Javier Sanchez and Dennys Gajdamaschko
  - for encouragement and discussion
- The people at FUNDEPS
  - Virginia Pedraza
- The people at TEC Costa Rica
  - Prof. Eddy Ramirez
- CLEI anonymous reviewers

`https://github.com/IE4OpenData/clei2019dialect`

Aitamurto, T., Chen, K., Cherif, A., Galli, J. S., and Santana, L. (2016).
Civic crowdanalytics: Making sense of crowdsourced civic input with big data tools.
In *Proceedings of the 20th International Academic Mindtrek Conference*, pages 86–94. ACM.

Bara, J., Weale, A., and Bicquelet, A. (2007).
Deliberative democracy and the analysis of parliamentary debate.
In *Workshop on Advanced Empirical Study of Deliberation*, pages 1–47.

Baud, R., Rassinoux, A.-M., and Scherrer, J.-R. (1992).
Natural language processing and semantical representation of medical texts.
Methods of information in medicine, 31(02):117–125.

Blodgett, S. L. and O'Connor, B. (2017).
Racial disparity in natural language processing: A case study of social media african-american english.
CoRR, abs/1707.00061.

Bogantes, D., Rodríguez, E., Arauco, A., Rodríguez, A., and Savary, A. (2016).
Towards lexical encoding of multi-word expressions in spanish dialects.
In *Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Carpuat, M. (2014).
Mixed language and code-switching in the canadian hansard.
In *Proceedings of the first workshop on computational approaches to code switching*, pages 107–115.

Ciobanu, A. M. and Dinu, L. P. (2016).

A computational perspective on the romanian dialects.
In *LREC*.

Dahlmeier, D. (2017).
On the challenges of translating nlp research into commercial products.
In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–96.

de Benito Moreno, C., Pueyo, J., and Fernández-Ordóñez, I. (2016).
Creating and designing a corpus of rural spanish.
In *Procceedings of the 13th Conference on Natural Language Processing (KONVENS)*, pages 78–83.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018).

Bert: Pre-training of deep bidirectional transformers for language understanding.
arXiv preprint arXiv:1810.04805.


Diab, M. and Habash, N. (2014).
Natural language processing of arabic and its dialects.
In *Tutorial at EMNLP*. ACL.


Ferrari, A., Donati, B., and Gnesi, S. (2017).
Detecting domain-specific ambiguities: an nlp approach based on wikipedia crawling and word embeddings.
In *2017 IEEE 25th International Requirements Engineering Conference Workshops (REW)*, pages 393–399. IEEE.


Ferraro, J. P., Daumé III, H., DuVall, S. L., Chapman, W. W., Harkema, H., and Haug, P. J. (2013).

Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation.
Journal of the American Medical Informatics Association, 20(5):931–939.


Fonseca, E. R., Rosa, J. L. G., and Aluísio, S. M. (2015).
Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese.
Journal of the Brazilian Computer Society, 21(1):2.


Ford, E., Carroll, J. A., Smith, H. E., Scott, D., and Cassell, J. A. (2016).
Extracting information from the text of electronic medical records to improve case detection: a systematic review.
Journal of the American Medical Informatics Association, 23(5):1007–1015.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017).
Portuguese word embeddings: evaluating on word analogies and natural language tasks.
arXiv preprint arXiv:1708.06025.

Hashemi, H. B. and Hwa, R. (2016).
An evaluation of parser robustness for ungrammatical sentences.
In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1765–1774.

Hendrycks, D. and Gimpel, K. (2016).
Gaussian error linear units (gelus).
arXiv preprint arXiv:1606.08415.

Ilie, C. (2004).

Insulting as (un) parliamentary practice in the british and swedish parliaments.
Cross-cultural perspectives on parliamentary discourse, 10:45.

Lei, Y. and Hansen, J. H. (2009).
Factor analysis-based information integration for arabic dialect identification.
In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4337–4340. IEEE.

Logeswaran, L. and Lee, H. (2018).
An efficient framework for learning sentence representations.
In *Proc. of ICLR2018*.

Logeswaran, L., Lee, H., and Radev, D. (2018).
Sentence ordering and coherence modeling using recurrent neural networks.
In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y. (2014).
Balanced corpus of contemporary written japanese.
Language resources and evaluation, 48(2):345–371.

Miller, T., Bethard, S., Amiri, H., and Savova, G. (2017).
Unsupervised domain adaptation for clinical negation detection.
In *BioNLP 2017*, pages 165–170.

Monroe, B. L. and Schrodt, P. A. (2008).
Introduction to the special issue: The statistical analysis of political text.
Political Analysis, 16(4):351–355.

Onyimadu, O., Nakata, K., Wilson, T., Macken, D., and Liu, K. (2013).

Towards sentiment analysis on parliamentary debates in hansard.
In *Joint international semantic technology conference*, pages 48–50. Springer.


Park, Y., Patwardhan, S., Visweswariah, K., and Gates, S. C. (2008).
An empirical analysis of word error rate and keyword error rate.
In *Ninth Annual Conference of the International Speech Communication Association*.


Peng, N., Yu, M., and Dredze, M. (2015).
An empirical study of chinese name matching and applications.

In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint*

Conference on Natural Language Processing (Volume 2: Short Papers), volume 2, pages 377–383.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018).
Deep contextualized word representations.
arXiv preprint arXiv:1802.05365.

Plank, B. (2016).
What to do about non-standard (or non-canonical) language in nlp.
In *In Proc. KOVENS 2016*.

Poole, K. D., Berson, M. J., and Levine, P. (2010).
On becoming a legislative aide: Enhancing civic engagement through a digital simulation.
Action in Teacher Education, 32(4):70–82.

Qu, L., Ferraro, G., Zhou, L., Hou, W., Schneider, N., and Baldwin, T. (2015).
Big data small data, in domain out-of domain, known word unknown word: The impact of word representations on sequence labelling tasks.
In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 83–93.

Radford, W. and Gallé, M. (2016).
Discriminating between similar languages in twitter using label propagation.
arXiv preprint arXiv:1607.05408.

Rasiah, P. (2010).
A framework for the systematic analysis of evasion in parliamentary discourse.
Journal of Pragmatics, 42(3):664–680.

Rasiah, P. et al. (2010).
Can the opposition effectively ensure government accountability in question time?: an empirical study.
Australasian Parliamentary Review, 25(1):166.

Rodríguez, C. F. (2011).
Cortesía e imagen en las preguntas orales del parlamento español.
Cultura, Lenguaje y Representación/Culture, Language and Representation, 9(9):53–79.

Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P., Morris, M., Yu, H., Duboue, P., Weng, W., Wilbur, W. J., et al. (2004).
Geneways: a system for extracting, analyzing, visualizing, and integrating molecular pathway data.
Journal of biomedical informatics, 37(1):43–53.

Sanchez-Perez, M. A., Markov, I., Gómez-Adorno, H., and Sidorov, G. (2017).
Comparison of character n-grams and lexical features on author, gender, and language variety identification on the same spanish news corpus.
In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 145–151. Springer.

Shaalan, K., Siddiqui, S., Alkhatib, M., and Monem, A. A. (2018).
Challenges in arabic natural language processing.
Computational Linguistics, Speech And Image Processing For Arabic Language, 4:59.

Shoufan, A. and Alameri, S. (2015).
Natural language processing for dialectical arabic: A survey.

In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 36–48.

Tatman, R. (2017).
Gender and dialect bias in YouTube's automatic captions.
In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017).
Attention is all you need.
In *Advances in neural information processing systems*, pages 5998–6008.

Wattam, S., Rayson, P., Alexander, M., and Anderson, J. (2014).

Experiences with parallelisation of an existing nlp pipeline: Tagging hansard.
In *LREC*, pages 4093–4096.


Whittle, J., Simm, W., Ferrario, M.-A., Frankova, K., Garton, L., Woodcock, A., Binner, J., Ariyatum, A., et al. (2010).
Voiceyourview: collecting real-time feedback on the design of public spaces.
In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 41–50. ACM.


Wong, T.-s., Gerdes, K., Leung, H., and Lee, J. (2017).
Quantitative comparative syntax on the cantonese-mandarin parallel dependency treebank.
In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 266–275.

- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016).
  Google's neural machine translation system: Bridging the gap between human and machine translation.
  arXiv preprint arXiv:1609.08144.
- Xu, F., Mingwen, W., and Li, M. (2016).
  Sentence-level dialects identification in the greater china region.
  International Journal on Natural Language Computing, 5:9–20.
- Xu, F., Wang, M., and Li, M. (2018).
  Building parallel monolingual gan chinese dialects corpus.
  In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.
- Zhang, Z.-s. (2017).

Dimensions of variation in written Chinese. Routledge.