TArC: Tunisian Arabish Corpus

First complete release

Elisa Gugliotta and Marco Dinarelli

Université de Grenoble Alpes (LIG and LIDILEM), Sapienza University of Rome

LREC 2022, June, Marseille

Overview

Corpus Usefulness

- A Computational Point of View:

Corpus Usefulness

- A Computational Point of View:
  - Normalized data availability for NLP research;

Corpus Usefulness

- A Computational Point of View:
  - Normalized data availability for NLP research;
  - Data annotation levels produced with reproducible and extensible methodology;

Corpus Usefulness

- A Computational Point of View:
  - Normalized data availability for NLP research;
  - Data annotation levels produced with reproducible and extensible methodology;
- A Linguistic Point of View:

Corpus Usefulness

- A Computational Point of View:
  - Normalized data availability for NLP research;
  - Data annotation levels produced with reproducible and extensible methodology;
- A Linguistic Point of View:
  - Provision of data for linguistic research;

Corpus Usefulness

- A Computational Point of View:
  - Normalized data availability for NLP research;
  - Data annotation levels produced with reproducible and extensible methodology;
- A Linguistic Point of View:
  - Provision of data for linguistic research;
  - Linguistic annotations and metadata to support multiple types of analysis;

The Corpus - Data Collection

- 'A Frequency Dictionary of Arabic', and in particular its 'Thematic Vocabulary List' (TVL) (Buckwalter and Parkinson, 2014).

The Corpus - Data Collection

- 'A Frequency Dictionary of Arabic', and in particular its 'Thematic Vocabulary List' (TVL) (Buckwalter and Parkinson, 2014).
- 'Loanword Typology Meaning List', which is a list of 1460 meanings (LTML) (Haspelmath and Tadmor, 2009).

The Corpus - Data Collection

- 'A Frequency Dictionary of Arabic', and in particular its 'Thematic Vocabulary List' (TVL) (Buckwalter and Parkinson, 2014).
- 'Loanword Typology Meaning List', which is a list of 1460 meanings (LTML) (Haspelmath and Tadmor, 2009).

The Corpus - Data Collection

- 'A Frequency Dictionary of Arabic', and in particular its 'Thematic Vocabulary List' (TVL) (Buckwalter and Parkinson, 2014).
- 'Loanword Typology Meaning List', which is a list of 1460 meanings (LTML) (Haspelmath and Tadmor, 2009).

Macro-Categories	Words Associated
1. Family *son, wedding, divorce*	w(e|i|ɛ)ld, 3(a|e)rs, (t|6)l(a|e)(9|q)
2. Clothing *dress, shoes, T-shirt*	r(o|ou)b(e|a), lebsa, sabat, sp(e|a)dri, m(a|e)r(i|y)oul

Table 1: Sample of the 15 thematic categories

The Corpus - Data Collection

- 'A Frequency Dictionary of Arabic', and in particular its 'Thematic Vocabulary List' (TVL) (Buckwalter and Parkinson, 2014).
- 'Loanword Typology Meaning List', which is a list of 1460 meanings (LTML) (Haspelmath and Tadmor, 2009).

Macro-Categories	Words Associated
1. Family *son, wedding, divorce*	w(e|i|ɛ)ld, 3(a|e)rs, (t|6)l(a|e)(9|q)
2. Clothing *dress, shoes, T-shirt*	r(o|ou)b(e|a), lebsa, sabat, sp(e|a)dri, m(a|e)r(i|y)oul

Table 1: Sample of the 15 thematic categories

	Sentences	*Words*	*Avg phrase len.*
Total	4,797	43,327	9.0
forum	755	11,909	15.8
social	3,162	16,056	5.1
blog	366	6,671	18.2
rap	514	8,691	16.9

Table 2: Statistics of our corpus

The Corpus - Semi-Automatic Annotation (1^{st} phase)

- TArC split in 7 blocks (of roughly 6,000 tokens each one).

The Corpus - Semi-Automatic Annotation (1^{st} phase)

- TArC split in 7 blocks (of roughly 6,000 tokens each one).
- Normalization in CODA Star (Habash et al., 2018) with an iterative procedure:

The Corpus - Semi-Automatic Annotation (1^{st} phase)

- TArC split in 7 blocks (of roughly 6,000 tokens each one).
- Normalization in CODA Star (Habash et al., 2018) with an iterative procedure:

  1. Annotating automatically a block of data with a model;

The Corpus - Semi-Automatic Annotation (1^{st} phase)

- TArC split in 7 blocks (of roughly 6,000 tokens each one).
- Normalization in CODA Star (Habash et al., 2018) with an iterative procedure:

1. Annotating automatically a block of data with a model;
2. Correcting manually the automatic annotation;

The Corpus - Semi-Automatic Annotation (1^{st} phase)

- TArC split in 7 blocks (of roughly 6,000 tokens each one).
- Normalization in CODA Star (Habash et al., 2018) with an iterative procedure:

1. Annotating automatically a block of data with a model;
2. Correcting manually the automatic annotation;
3. Adding the new annotated block of data to the training data of the model;

The Corpus - Semi-Automatic Annotation (1^{st} phase)

- TArC split in 7 blocks (of roughly 6,000 tokens each one).
- Normalization in CODA Star (Habash et al., 2018) with an iterative procedure:

1. Annotating automatically a block of data with a model;
2. Correcting manually the automatic annotation;
3. Adding the new annotated block of data to the training data of the model;
4. Training a new model;

The Corpus - Semi-Automatic Annotation (1^{st} phase)

- TArC split in 7 blocks (of roughly 6,000 tokens each one).
- Normalization in CODA Star (Habash et al., 2018) with an iterative procedure:

1. Annotating automatically a block of data with a model;
2. Correcting manually the automatic annotation;
3. Adding the new annotated block of data to the training data of the model;
4. Training a new model;
5. Restarting from step 1 with a new block of data.

The Corpus - Semi-Automatic Annotation (1^{st} phase)

- TArC split in 7 blocks (of roughly 6,000 tokens each one).
- Normalization in CODA Star (Habash et al., 2018) with an iterative procedure:

1. Annotating automatically a block of data with a model;
2. Correcting manually the automatic annotation;
3. Adding the new annotated block of data to the training data of the model;
4. Training a new model;
5. Restarting from step 1 with a new block of data.

- Last step accuracy of 65%.

The Corpus - Semi-Automatic Annotation (2^{nd} phase)

Use of the MADAR corpus (Bouamor et al., 2018).

The Corpus - Semi-Automatic Annotation (2^{nd} phase)

Use of the MADAR corpus (Bouamor et al., 2018).

1. Classification (*arabizi*, *foreign*, *emotag*);

The Corpus - Semi-Automatic Annotation (2^{nd} phase)

Use of the MADAR corpus (Bouamor et al., 2018).

1. Classification (*arabizi*, *foreign*, *emotag*);
2. Transliteration (CODA*);

The Corpus - Semi-Automatic Annotation (2^{nd} phase)

Use of the MADAR corpus (Bouamor et al., 2018).

1. Classification (*arabizi*, *foreign*, *emotag*);
2. Transliteration (CODA*);
3. Tokenization (Word > Morphemes);

The Corpus - Semi-Automatic Annotation (2^{nd} phase)

Use of the MADAR corpus (Bouamor et al., 2018).

1. Classification (*arabizi*, *foreign*, *emotag*);
2. Transliteration (CODA*);
3. Tokenization (Word > Morphemes);
4. POS-tagging (PATB (Maamouri et al., 2004));

The Corpus - Semi-Automatic Annotation (2^{nd} phase)

Use of the MADAR corpus (Bouamor et al., 2018).

1. Classification (*arabizi*, *foreign*, *emotag*);
2. Transliteration (CODA*);
3. Tokenization (Word > Morphemes);
4. POS-tagging (PATB (Maamouri et al., 2004));
5. Lemmatization (CODA*).

The Corpus - Semi-Automatic Annotation (2^{nd} phase)

Use of the MADAR corpus (Bouamor et al., 2018).

1. Classification (*arabizi*, *foreign*, *emotag*);
2. Transliteration (CODA*);
3. Tokenization (Word > Morphemes);
4. POS-tagging (PATB (Maamouri et al., 2004));
5. Lemmatization (CODA*).

The Corpus - Semi-Automatic Annotation (2^{nd} phase)

Use of the MADAR corpus (Bouamor et al., 2018).

1. Classification (*arabizi*, *foreign*, *emotag*);
2. Transliteration (CODA*);
3. Tokenization (Word > Morphemes);
4. POS-tagging (PATB (Maamouri et al., 2004));
5. Lemmatization (CODA*).

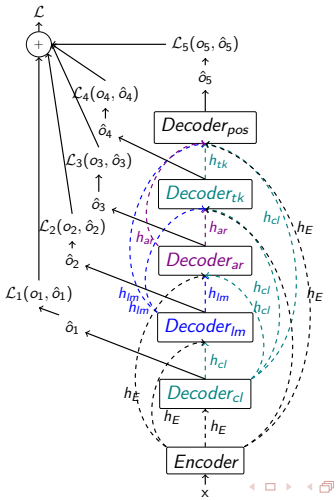
CODA	**Tokeniz.**	**POS**	**Lemma**
انا	انا	PRON_1S	هو
بعد	بعد	ADV	بعد
ma	foreign	foreign	foreign
grossesse	foreign	foreign	foreign
حوايجي	حوايج+ي	NOUN+ POSS_PRON_1S	حوايج
ال	ال	DET	ال
قدم	قدم	ADJ	قديم
ال	ال	DET	ال
كلّهم	كلّ+هم	NOUN_QUANT+ PRON_3P	كلّ
ولّاوا	ولّاوا	PV-PVSUFF_ SUBJ:3P	ولّى
motivation	foreign	foreign	foreign

Table 3: TArC Annotation Levels

The Neural Multi-Task Architecture

- x = input in Arabizi;

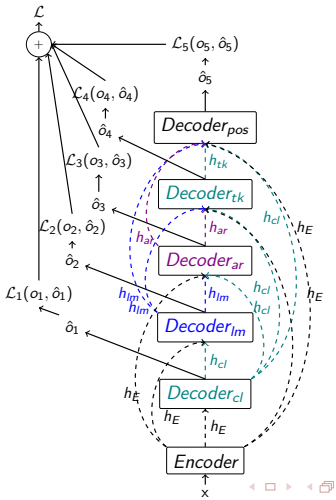
Figure 1: The Architecture Schema



The Neural Multi-Task Architecture

- x = input in Arabizi;
- An encoder to convert x into context-aware repr.;

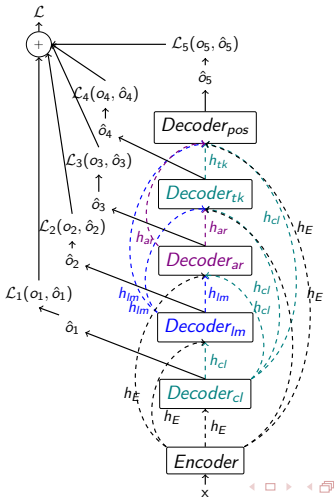
Figure 1: The Architecture Schema



The Neural Multi-Task Architecture

- x = input in Arabizi;
- An encoder to convert x into context-aware repr.;
- 5 Decoders (one for each annotation level);

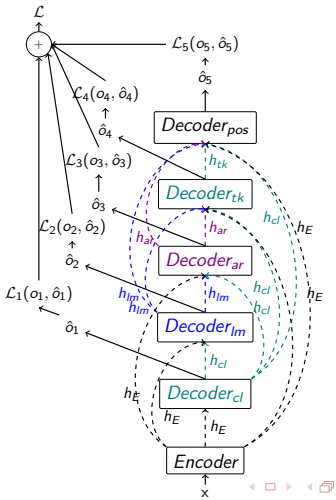
Figure 1: The Architecture Schema



The Neural Multi-Task Architecture

- x = input in Arabizi;
- An encoder to convert x into context-aware repr.;
- 5 Decoders (one for each annotation level);
- Attention mechanism to pass previous hidden states ($h_{i-j_{<i}}$) to the modules;

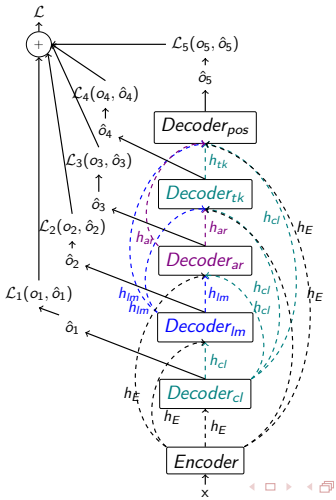
Figure 1: The Architecture Schema



The Neural Multi-Task Architecture

- x = input in Arabizi;
- An encoder to convert x into context-aware repr.;
- 5 Decoders (one for each annotation level);
- Attention mechanism to pass previous hidden states ($h_{i-j_{<i}}$) to the modules;
- Single task loss computation (i.e. $\mathcal{L}_i(o_i, \hat{o}_i)$);

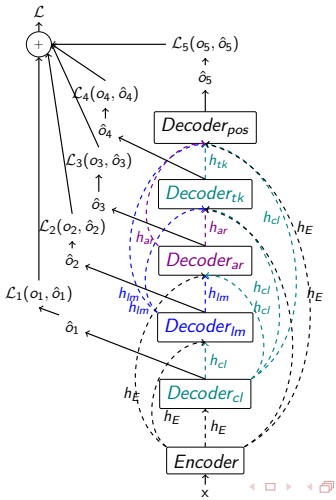
Figure 1: The Architecture Schema



The Neural Multi-Task Architecture

- x = input in Arabizi;
- An encoder to convert x into context-aware repr.;
- 5 Decoders (one for each annotation level);
- Attention mechanism to pass previous hidden states ($h_{i-j_{<i}}$) to the modules;
- Single task loss computation (i.e. $\mathcal{L}_i(o_i, \hat{o}_i)$);
- End-to-end learning of the whole architecture: $\mathcal{L} = \sum_{i=1}^{5} \mathcal{L}_i(o_i, \hat{o}_i)$.

Figure 1: The Architecture Schema



Experiments - Before Lemmatization

Summary of results, in terms of accuracy, obtained on the TArC data at the different steps of the iterative procedure for semi-automatic annotation of the corpus. The tasks are indicated with **CI** for classification, **Ar** for Arabic script encoding, **Tk** for tokenization, and **POS** for POS tagging. $^{(*)}$ indicates results obtained with the MADAR data translated into Arabizi.

Step	Train. tokens	Tasks (Accuracy) CI	Ar	Tk	POS
Corpus: MADAR					
Step0	12,391	99.83%	-	88.83%	72.71%
$\text{Step0}^{(*)}_{\text{complete}}$	12,391	99.58%	76.77%	74.83%	67.59%
Corpus: MADAR+TArC					
Step1	17,261 (4,870)	92.69%	-	77.66%	59.56%
Step2	22,173 (9,780)	97.21%	-	87.53%	74.30%
Step3	27,270 (14,870)	96.69%	-	91.47%	76.38%
Corpus: TArC					
Step4	22,150	96.83%	75.30%	73.38%	69.76%
Step5	27,435	97.17%	75.08%	73.07%	66.24%
$\text{Step4}_{\text{smart-init}}$	22,150	95.91%	76.55%	74.96%	72.57%
$\text{Step5}_{\text{smart-init}}$	27,435	97.08%	77.83%	75.69%	69.76%
Corpus: $\text{MADAR}_{\text{Arabizi}}$+TArC					
$\text{Step4}_{\text{concat}}{}^{(*)}$	34,541 (22,150)	96.59%	78.94%	77.38%	74.54%
$\text{Step4}_{\text{reloaded}}{}^{(*)}$	34,541 (22,150)	96.38%	79.72%	77.88%	73.69%
$\text{Step6}_{\text{concat}}{}^{(*)}$	46,197 (33,806)	96.45%	79.97%	77.81%	70.33%
$\text{Step6}_{\text{concat}}{}^{(*)}$ **fix**	46,197 (33,806)	97.63%	83.29%	81.94%	81.02%
$\text{Final-Step}_{\text{concat}}{}^{(*)}$ lstm	42,895 (30,504)	98.56%	82.98%	81.84%	82.84%
$\text{Final-Step}_{\text{concat}}{}^{(*)}$ transformer	42,895 (30,504)	95.99%	75.37%	74.34%	71.30%
$\text{Final-Step}_{\text{concat}}{}^{(*)}$ input:Ar lstm	42,895 (30,504)	98.67%	-	96.78%	86.31%
$\text{Final-Step}_{\text{concat}}{}^{(*)}$ input:Ar transformer	42,895 (30,504)	99.95%	-	95.93%	82.49%

Lemmatization

- Tool of fundamental importance both for linguistic analysis and automatic data processing (Zalmout and Habash, 2019);

Lemmatization

- Tool of fundamental importance both for linguistic analysis and automatic data processing (Zalmout and Habash, 2019);
- TArC lemmas are also encoded in CODA* (Habash et al., 2018);

Lemmatization

- Tool of fundamental importance both for linguistic analysis and automatic data processing (Zalmout and Habash, 2019);
- TArC lemmas are also encoded in CODA* (Habash et al., 2018);
- Employment of the semi-automatic annotation procedure used for the other levels (5) (Gugliotta et al., 2020);

Lemmatization

- Tool of fundamental importance both for linguistic analysis and automatic data processing (Zalmout and Habash, 2019);
- TArC lemmas are also encoded in CODA* (Habash et al., 2018);
- Employment of the semi-automatic annotation procedure used for the other levels (5) (Gugliotta et al., 2020);
- Procedure bootstrapping by manually lemmatizing a first block of TArC;

Lemmatization

- Tool of fundamental importance both for linguistic analysis and automatic data processing (Zalmout and Habash, 2019);
- TArC lemmas are also encoded in CODA* (Habash et al., 2018);
- Employment of the semi-automatic annotation procedure used for the other levels (5) (Gugliotta et al., 2020);
- Procedure bootstrapping by manually lemmatizing a first block of TArC;
- Employment of MADAR (Bouamor et al., 2018) semi-automatically lemmatized.

Lemmatization - Summary of Experiments Results

Step	Train. tokens	Tasks (Accuracy)				
		Class	Arabic	Token	PoS	Lemma
Corpus: MADAR$_{Arabizi}$+TArC						
Step1	17,509 (5118)	98.61	73.61	73.15	73.92	72.22
Step2	22,272 (9,881)	97.33	79.10	77.53	78.82	75.14
Step3	27,399 (15,008)	98.31	80.69	79.81	80.38	79.00
Step4	33,069 (20,678)	99.13	81.77	80.94	82.30	80.36
Step5	38,681 (26,290)	98.72	85.79	84.89	85.44	83.69
Step6	44,792 (32,401)	97.13	85.96	84.81	83.11	84.38
Final Step global-split	42,559 (30,168)	97.14	82.34	81.45	80.95	80.48
Final Step genre-split	42,559 (30,168)	98.47	82.93	81.77	80.33	81.40
Step	**Train. tokens**	**Class**	**Lemma**	**Arabic**	**Token**	**PoS**
Final Step 2xlstm	42,559 (30,168)	98.42	81.81	82.65	81.58	81.60
Final Step 3xtransformer	42,559 (30,168)	96.48	68.89	69.72	68.18	68.37
Final Step 2xlstm input:Ar	42,559 (30,168)	98.77	92.40	-	96.74	85.90
Final Step 3xtransformer input:Ar	42,559 (30,168)	96.91	83.10	-	93.43	74.09

Conclusion and Future Works

- We aimed at:

Conclusion and Future Works

- We aimed at:
  - Providing a response to the lack of tools to support both NLP and linguistic research on Tunisian Arabic;

Conclusion and Future Works

- We aimed at:
  - Providing a response to the lack of tools to support both NLP and linguistic research on Tunisian Arabic;
  - Building a corpus suitable for various type of analyses;

Conclusion and Future Works

- We aimed at:
  - Providing a response to the lack of tools to support both NLP and linguistic research on Tunisian Arabic;
  - Building a corpus suitable for various type of analyses;
- Next objective is:

Conclusion and Future Works

- We aimed at:
  - Providing a response to the lack of tools to support both NLP and linguistic research on Tunisian Arabic;
  - Building a corpus suitable for various type of analyses;
- Next objective is:
  - TArC extension with the last MADAR release (Eryani et al., 2020);

Conclusion and Future Works

- We aimed at:
  - Providing a response to the lack of tools to support both NLP and linguistic research on Tunisian Arabic;
  - Building a corpus suitable for various type of analyses;
- Next objective is:
  - TArC extension with the last MADAR release (Eryani et al., 2020);

Figure 2: Architecture

Conclusion and Future Works

- We aimed at:
  - Providing a response to the lack of tools to support both NLP and linguistic research on Tunisian Arabic;
  - Building a corpus suitable for various type of analyses;
- Next objective is:
  - TArC extension with the last MADAR release (Eryani et al., 2020);

Figure 2: Architecture

Figure 3: TArC

Conclusion and Future Works

- We aimed at:
  - Providing a response to the lack of tools to support both NLP and linguistic research on Tunisian Arabic;
  - Building a corpus suitable for various type of analyses;
- Next objective is:
  - TArC extension with the last MADAR release (Eryani et al., 2020);

Figure 2: Architecture

Thanks for your attention!

Figure 3: TArC

References I

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. The madar arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Tim Buckwalter and Dilworth Parkinson. *A frequency dictionary of Arabic: Core vocabulary for learners*. Routledge, 2014.

Fadhl Eryani, Nizar Habash, Houda Bouamor, and Salam Khalifa. A spelling correction corpus for multiple arabic dialects. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4130–4138, 2020.

References II

Elisa Gugliotta, Marco Dinarelli, and Olivier Kraif. Multi-task sequence prediction for Tunisian Arabizi multi-level annotation. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 178–191, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.wanlp-1.16`.

Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, et al. Unified guidelines and resources for arabic dialect orthography. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Martin Haspelmath and Uri Tadmor. The loanword typology meaning list: Electronic databases of 29 languages. *A collaborative project coordinated by the Max Planck Institute for Evolutionary Anthropology, Department of Linguistics*, 2009.

References III

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. The Penn Arabic Treebank: Building a large-scale annotated Arabic corpus. In *NEMLAR conference on Arabic language resources and tools*, volume 27, pages 466–467. Cairo, 2004.

Nasser Zalmout and Nizar Habash. Joint diacritization, lemmatization, normalization, and fine-grained morphological tagging. *arXiv preprint arXiv:1910.02267*, 2019.