Building a Language Model for Tunisian Dialect

Nourchene Ferchichi, Eya Rhouma, Mohamed Salam Jedidi, Amine Kerkeni

InstaDeep™

ABSTRACT

The Tunisian dialect is different from the standard Arabic language due to its complicated written forms. We add to that the nonexistence of a sufficient large Tunisian corporus.

Our challenge is to collect a Tunisian dataset and create a model that can understand this dialect by testing it on the "Multi-Turn Response Selection" task. We propose to use the "Deep Attention Matching" (DAM) model [1], a recently introduced model based entirely on "Attention" and inspired by the "Transformer" model [2].

Experiments on our collected and pre-processed 42K Tunisian dataset show that the proposed solution was successful. Word Embedding succeeded in tackling all the dialect complexity and the model reached a Recall R10@5 of more than 0,95 on the final tests.

TUNISIAN DIALECT DATASET

Tunisian Dialect Challenges

The Tunisian dialect is very complicated when it comes to understanding its written forms. What distinguishes this dialect is the mixture of Latin and Arabic alphabet, and the presence of numbers coding letters. We add to that the absence of spelling rules, and the presence of foreign languages with misspelling in them.

We notice also the lack of sufficiently large datasets. Infact, very few freely available Tunisian corpora can be found. Moreover, the already existing datasets are either designed for sentiment analysis or linguistic experiments.

Dataset Statistics

We provide a new Tunisian corpus, OoredooTn dataset, for "Multi-turn Response Selection" task. The dataset contains more than 42K conversations about technical information extracted from the "Ooredoo Tunisie", a Tunisian telecommunications company, Facebook page comments and replies. It is on a less scale than standard language datasets (Ubuntu an Douban datasets) for solving dialogue problems (Table 1). However, it has an average of more than 2 turns each. Adding to that, each conversation in our dataset includes long utterances.

	Ubuntu [3]			Douban [4]			OoredooTN		
	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
Context response pairs	1M	500K	500K	1M	50K	10K	**20K**	**11.3K**	**11.29K**
Candidates per context	2	10	10	2	2	10	**2**	**10**	**10**
Avg turns per context	10.13	10.11	10.11	6.69	6.75	6.45	**2.25**	**2.10**	**2.06**
Avg turns per words	11.35	11.34	11.37	18.56	18.50	20.74	**16.73**	**17.05**	**15.77**

Table 1: Statistics of the OoredooTn dataset.

MODEL ARCHITECTURE

The "Deep Attention Matching" (DAM) model [1] is composed of four modules: The Word Embedding, the Representation, the Matching, and the Aggregation modules (Figure 2).

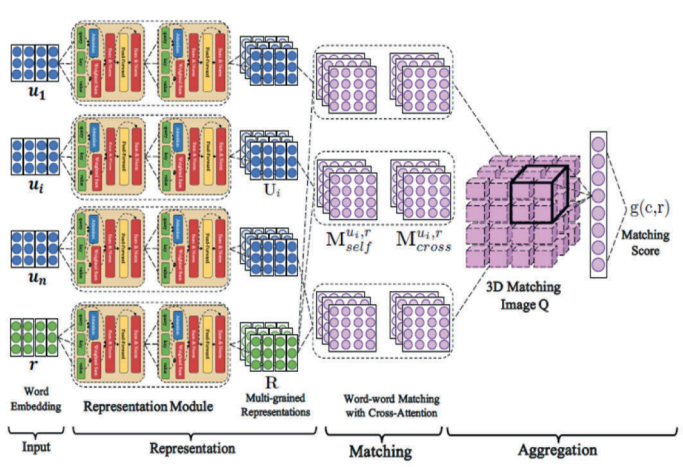

Figure 2: The Deep Attention Matching Model (DAM) architecture [1].

DAM takes each word of an utterance in context or response and hierarchically enriches its representation with successive levels of "Attention" modules. This gradually produces sophisticated segment representations surrounding the word from one level to another. In this way, each utterance in context and response is matched based on segment pairs at different levels. Therefore, DAM captures matching information between the context and the response from word-level to sentence-level, until arriving to the context-response-level.

EXPERIMENTAL RESULTS

Word Embedding

The Embedding module solved one side of the Tunisian dialect challenges, which is its language complexity. In fact, this module detected the presence of foreign languages and added a translation meaning to them (Figure 3).

It also extracted unfixed spelling to the same words, and detected the use of both Latin and Arabic alphabet. Finally it pointed the misspelling when it came to foreign languages (Figure 4).

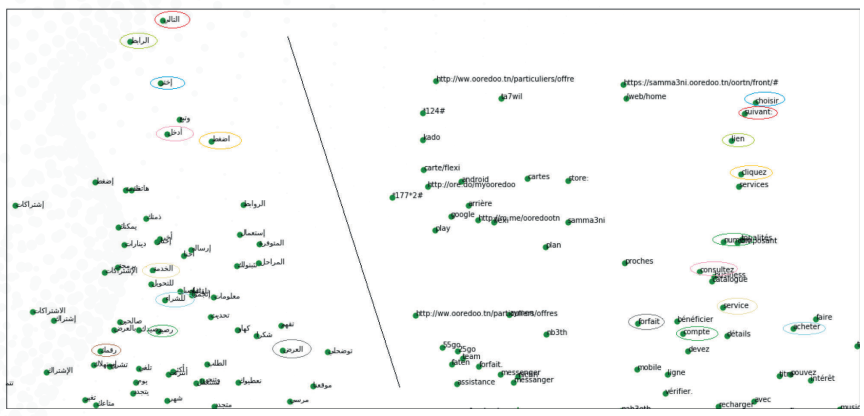

Figure 3: Word embedding representation.

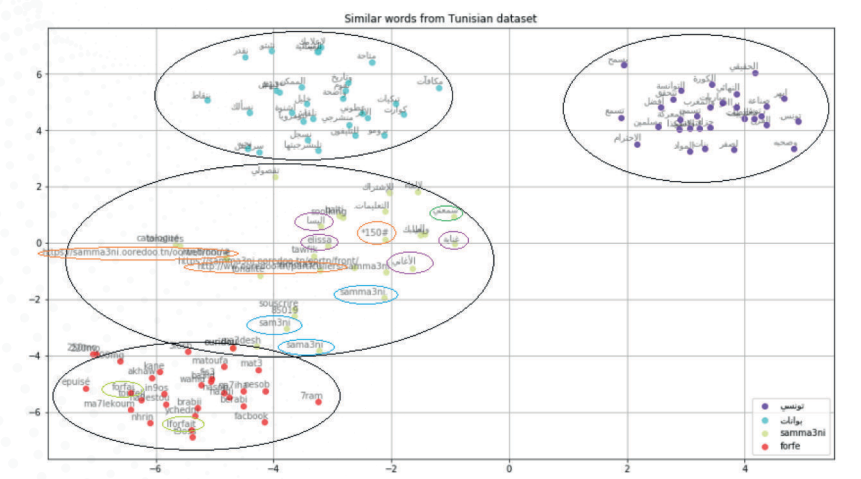

Figure 4: Word embedding representation for four extracted clusters.

Training, validation and test

DAM shows good performance at matching responses with short context with only 2 utterances (Left Figure 5). It can still deal with long context length with more than 6 turns. DAM reacts well when it comes to long turns with more then 10 words per turn (Right Figure 5). Unfortunately, its performance on short utterances with less than 10 words, is lower. This is because the shorter the utterance is, the fewer information it contains, and the more difficult it is for selecting the next utterance.

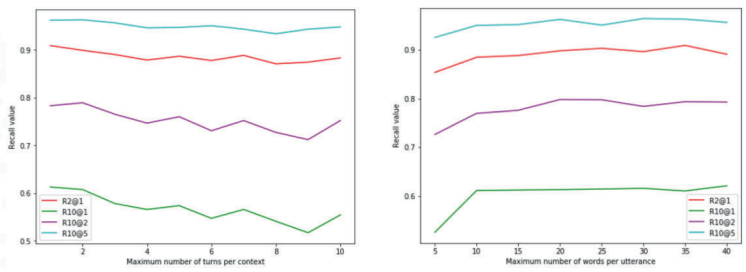

Figure 5: The model performances on OoredooTn dataset.

We deliver test results that reveal the model performances on the "Multi-turn context response selection" task. The overall results outperformed Douban dataset result and are close to the Ubuntu dataset.The results show the effectiveness of the DAM model with more than 0,95 for the recall R10@5 value.

	R2@1	R10@1	R10@2	R10@5
Ubuntu [3]	0.938	0.767	0.874	0.969
Douban [4]	- - -	0.254	0.410	0.757
OoredooTN	**0.892**	**0.588**	**0.770**	**0.952**

Figure 3: Best Recall result on OoredooTn dataset.

Conclusion

We collected a Tunisian dialect dataset that will be valuable for researchers working in the field of Tunisian language processing models. We applied the DAM model on the "Multi-turn Response selection" task and obtained successful results that can serve as comparative examples for future improved works.

REFERENCES

[1] Daxiang Dong Yi Liu Ying Chen Wayne Xin Zhaoy Dianhai Yu Xiangyang Zhou, Lu Li and Hua Wu.
Multi-turn response selection for chatbots withh deep attention matching network. July 2018.

[2] Google Brain. Attention is all you need. Decembre 2017.

[3] Iulian V.Serban† Ryan Lowe, Nissan Pow and Joelle Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. 2015.

[4] Chen Xing Zhoujun Li Ming Zhou Yu Wu, Wei Wu. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. 2017.