Introduction

- Given the unlimited access to the internet, many languages of users written spontaneously which called Arabic dialect (AD) are presented on social networks (SN)
- The need to process AD written on the SN allows to facilitate several NLP tasks such as opinions analysis

=> Sentiment Analysis of the spoken Arabic Tunisian dialect (TD)

Problematic:
- * TD written on SN are coupled by other languages/Foreign Languages, SMS language...
- * Lack of resources for TD (parallel corpus, annotated corpus...)

=> Several solutions to adapt for the sentiment analysis of the TD:
1. Create an analysis model specified for the TD
2. Create a multi-idiom model applicable on TD
3. Create a standard template for a standard language.

Objectives:

=> Sentiment Analysis of TD using a standard model:
 1. Identify the TD on SN
 2. Translate the identified TD into a standard language:
Modern Standard Arabic (MSA)
3. Apply a standard sentiment analysis model on the translated TD.

Overview

- Several approaches are used for dialect identification:
  - Classical approaches:
    - (Kchau et al., 2019)
    - (MADAR Shared Task 2019)
  - Deep learning approaches:
    - (Issa et al., 2021)
    - (NADI Shared Task 2020)
- Several approaches have been applied for Machine translation of TD:
  - Linguistic (Hamdi et al., 2013)
  - Statistical (Kchau et al., 2020)
  - No works dealing with Neural Machine Translation until now besides our work on the translation of Tunisian transcriptions (Abida et al., 2022)

Resources for Tunisian Dialect Identification

- Corpus for Dialect Identification: contains 96k annotated comments with 3 classes (Tun, MSA, Other), this corpus is collected from:
  1. (Kchau et al., 2020): Parallel corpus containing 32k parallel sentences TD-MSA built using an augmentation method applied on:
     - MADAR: Includes 1.8k parallel sentences in travel domain of 25 Arabic dialects and the MSA,
     - PADIC: Contains 6.4k parallel sentences from everyday life and television programs in the Maghreb dialects, Levant dialects and MSA,
     - Tunisian constitution: Includes 500 parallel sentences,
     - 900 TD comments manually translated by native speakers into MSA;
  2. NADI corpus: An annotated corpus at country level,

- Tunisian Arabic dialect identification (TADID) model:

  Traditional approaches:

  - Transfer learning approaches:

  - Multi-idiom-Arabic-BERT => TADID model is developed using the Multi-idiom-Arabic-BERT

Tunisian Dialect identification resources(2)

- Multi-idiom-Arabic-BERT = 8.82%

Resources for TD-MSA Translation

- TD-MSA parallel corpus:
  - 32k parallel sentences (Kchau et al., 2020)
  - Social media corpus: 2k comments predicted using TADID model and manually translated by native speakers,

- MAGES: Modern standard Arabic texts GEnration tool from Social media

- MAGES is a tool that combines the developed TADID model and TD-MSA-NMT model. Given a corpus taken from social networks, TADID model makes it possible to identify the AD and translates the AD comments to the MSA and leaves the comments written in MSA intact.

- To evaluate the MAGES tool, another test set was used: It contains 1408 comments: 500 parallel sentences TD-MSA used in (Kchau et al., 2020) and 468 comments in other languages.

- From the 1406 comments, MAGES generates 444 sentences in MSA among 500 MSA comments, i.e. an accuracy of 89%. It has correctly identified 410 TD comments.

- The tagged comments with the MSA class are passed to the output of the system, and the identified TD comments are passed to NMT model.

MAGES evaluation on the application of sentiment analysis

- The main objective of MAGES tool is to:
  - Facilitate the creation of parallel corpus,
  - Allow the application of MSA linguistic resources such as sentiment analysis

- Effect of MAGES on sentiment analysis of dialectal text content in social networks:
  - The 1406 comments of the test set are tagged by 3 classes (Positive, Negative, Neutral)

  - Two pre-trained models CAMeLBERT (Issa et al., 2021) are used in order to compare sentiment analysis of TD comments and their correspondence in MSA:
    - CAMeLBERT-AD for sentiment analysis of TD (or Mages system input)
    - CAMeBERT-MSA model for sentiment analysis of MSA texts (Mages system output)

  => Results show that the approach of standardization of dialectal content is better than that of independent treatment of Arabic dialects,

Conclusion

- Parallel corpus containing 64k parallel sentences is created in which 2k parallel sentences TD-MSA are manually built and made available to researchers.
- An identification model for TD and MSA from a corpus scraped from social networks is proposed.
- A model to standardize the written text TD texts in social networks in order to facilitate the computational analysis of poorly endowed languages is proposed in this work.
- An MSA text geneation tool (MAGES) is created in order to develop a sentiment analysis model for TD.

Future work:

- Introduce the written comments in Arabic: Arabic dialect written in Latin script.
- Exploit other advanced pretraining methods, in order to translate TD into a foreign language like English or French.
- Investigate the effectiveness of the proposed techniques on other Arabic dialects.

Bibliographical References