EMO Project

Regresión de intensidad emocional cross-lingüe (EMO)

Pablo García Quintas

Introducción:

La tarea de regresión de la intensidad de las emociones intenta determinar la cantidad de emoción en un texto breve, como el texto de un tweet.

Limitación de la tarea: Falta de datos etiquetados en español y en inglés. Por ello tenemos dos estrategias para abordar el problema:

- La primera estrategia implica utilizar la traducción automática para traducir el conjunto de datos de evaluación en español al inglés.
- La segunda estrategia consiste en utilizar la traducción automática para convertir el conjunto de entrenamiento y desarrollo en inglés al español.

¿Qué herramientas se han usado?

- Github
- Spyder -> Python
- Para los datos de entrenamiento → Onlinedoctranslator
- Para los datos de evaluación → LibreTranslate

Código

He editado el código de manera que al ejecutarlo en la aplicación Spyder, se nos van pidiendo por pantalla las especificaciones necesarias:

- Emoción a evaluar (anger, joy, sadness o fear)
- Idioma de los datos del Train (castellano o inglés)
- Idioma de los datos del Test (castellano o inglés)
- Opciones adicionales: ngrams, char_ngrams, etc.

```
if __name__ == '__main__':
    print("Elige de que emoción se evaluará entre las que hay: anger, joy, sadness, fear")
    emocion = input()

    print("Elige idioma del train/dev data: ingles->1 o castellano->2")
    idiomaTrain = input()
    if idiomaTrain == "1":
        aux1 = "../dataset/en"
    else:
        aux1 = "../dataset/Ingles -> Español"

    print("Elige idioma del test data: ingles->1 o castellano->2")
    idiomaTrain2 = input()
    if idiomaTrain2 == "1":
        aux2 = "../dataset/Español -> Ingles"
    else:
        aux2 = "../dataset/es"

    aux3 = "1"
    opciones = []
    while aux3 !="0":
        print("Elige las opciones opcionales: ngrams, char_ngrams, NRC_sent, NRC_emo o
NRC_hash:")
        opciones.append(input())
        print("Seguir añadiendo->1; Salir->0")
        aux3 = input()

    parser = argparse.ArgumentParser()
    parser.add_argument('-sd', '--src_dataset', default= aux1)
    parser.add_argument('-td', '--trg_dataset', default= aux2)
    parser.add_argument('-emo', '--emotion', default=emocion)
    parser.add_argument('-f', '--features', nargs='+', default=opciones)

    args = parser.parse_args()
```

Resultados sin aplicar estrategias

TEST	ANGER	JOY	SADNESS	FEAR
NGRAMS	0,06	0,23	-0,03	0
CHAR_NGRAMS	0,13	0,21	-0,11	-0,01
NGRAMS-NRC_EMO	0,06	0,17	0,02	-0,02
CHAR_NGRAMS-NRC-EMO	0,13	0,2	-0,08	-0,01
NGRAMS-NRC_SENT	0,03	0,15	-0,05	0,02
CHAR_NGRAMS-NRC_SENT	0,07	0,2	-0,1	-0,02
NGRAMS-NRC_HASH	0	0,09	0,04	0,07
CHAR_NGRAMS-NRC_HASH	0,04	0,21	-0,03	0,02
MEDIAS	0,065	0,1825	-0,0425	0,00625

Como se puede ver los resultados son muy malos, prácticamente dan de media 0 en todas las emociones salvo en Joy.

Estrategia 1 - Traducir los datos de evaluación a español

Vemos que en comparación a no usar estrategias, los resultados mejoran significativamente. En las emociones anger y joy pasamos de tener de media 0 a tener casi un 0,3

TEST	ANGER	JOY	SADNESS	FEAR
NGRAMS	0,24	0,16	-0,15	0,13
CHAR_NGRAMS	0,21	0,27	-0,19	0,08
NGRAMS-NRC_EMO	0,31	0,29	-0,06	0,29
CHAR_NGRAMS-NRC-EMO	0,25	0,33	-0,15	0,15
NGRAMS-NRC_SENT	0,31	0,31	0,09	0,24
CHAR_NGRAMS-NRC_SENT	0,25	0,34	-0,1	0,12
NGRAMS-NRC_HASH	0,31	0,28	0,04	0,28
CHAR_NGRAMS-NRC_HASH	0,26	0,31	-0,11	0,16
MEDIAS	0,2675	0,28625	-0,07875	0,18125

Estrategia 2 - Traducir los datos de entrenamiento a inglés

Vemos que en este caso, la estrategia 2, también tiene mejores valores de media que la opción de no usar estrategias

TEST	ANGER	JOY	SADNESS	FEAR
NGRAMS	0,11	0,14	-0,07	0,08
CHAR_NGRAMS	0,18	0,26	0,04	0,13
NGRAMS-NRC_EMO	0,11	0,13	-0,08	0,07
CHAR_NGRAMS-NRC-EMO	0,18	0,26	0,03	0,12
NGRAMS-NRC_SENT	0,15	0,16	-0,03	0,11
CHAR_NGRAMS-NRC_SENT	0,21	0,27	0,05	0,13
NGRAMS-NRC_HASH	0,08	0,13	-0,07	0,1
CHAR_NGRAMS-NRC_HASH	0,18	0,26	0,03	0,13
MEDIAS	0,15	0,20125	-0,0125	0,10875

Conclusiones

Aunque es cierto que ambas estrategias mejoran los resultados básicos, nos decantaríamos por la Estrategia 1. Es decir, traducir los datos de evaluación al idioma de los datos de entrenamiento, generalmente será la mejor opción siempre y cuando las traducciones sean consistentes y estén realizadas lo mejor posible. No hay que olvidar que estamos usando traductores automáticos y las traducciones NO son perfectas.

Muchas Gracias