Marc Bracons Cucó

Deep Learning per a l'anàlisi de textos
PRA2

1. Elecció de Datasets (25%)

A partir de la selecció dels datasets per a la PRA2, que s'usen tant a la secció 1 "Traducció Automàtica (TA)" com a la secció 2 "Detecció de NER i NEL" contesta les següents preguntes:

1. Descriu els datasets escollits. Indica l'origen, la quantitat de dades que contenen, la tipologia d'aquestes i justifica l'elecció.

Per a la secció 1, centrada en la traducció automàtica he escollit el corpus TED2020 (ca-en)[1], aquest conté aproximadament 4000 xerrades TED I TED-X fent un total de 52.696 frases traduïdes per voluntaris.

S'ha organitzat les dades en dos blocs de text paral·lels: un amb les frases en anglès i l'altre amb les corresponents en català. Cada bloc està organitzat de manera sincronitzada perquè la línia **n** de l'un s'ajusti a l'altre, mantenint les relacions entre frases. Aquesta estructura facilita tant la càrrega com l'alienació automàtica durant l'entrenament del model.

He triat aquest corpus perquè combina un volum suficient per a l'entrenament de models profunds i a la vegada és prou petit com per fer entrenament en un exercici on l'objectiu no és tant el resultat com els coneixements teòrics.

Per altra banda, per a la tasca de reconeixement i enllaç d'entitats (secció 2) he triat el conjunt CoNLL-2003 [2]. Està format per gairebé 15000 frases per entrenament i unes 7000 per validació, totes anotades per a quatre categories d'entitats: Location, Organizations, Persons i Miscellaneous entities. Al igual que en el cas anterior s'ha triat aquest joc de dades a causa del seu volum de registres, que permet, amb la seva varietat permet entrenar un model.

[1] OPUS – CORPORA. TED 2020. Juliol del 2020.
https://opus.nlpl.eu/TED2020/en&ca/v1/TED2020
[2] CoNLL003 (English-version). https://www.kaggle.com/datasets/alaakhaled/conll003-englishversion

2. Explica els problemes que has trobat a les dades i les activitats de neteja i preprocessament que has realitzat en preparar les dades per a la pràctica de TA i la de NER.

En el joc de dades de la secció 1, les dades tenien signes de puntuació i caràcters en majúscules. S'ha aplicat un preprocessament per eliminar la puntuació i convertir tots els caracàcters a minúscules, obtenint així una representació més homogènia i apta per al model. A més, he fet visualitzacions amb histogrames per avaluar la distribució de longituds de les frases, gràcies a les quals he pogut veure que la majoria són curtes, generalment per sota dels 50 tokens.

En canvi, en la secció 2, les dades inicialment s'han hagut d'adaptar a format spaCy. Per fer-ho, s'ha fet una conversió del format original CoNLL a l'estructura interna de spaCy, generant arxius (.spacy) adequats per a l'entrenament.

Marc Bracons Cucó

2. Traducció automàtica (50%)

Per a cadascun dels models de traducció automàtica creats, respon les preguntes següents:

1. Explica com has definit el primer model encoder-decoder i com has triat els paràmetres que s'hi fan servir. A més, explica quins valors dels paràmetres haguessin estat més apropiats, depenent de les dades de l'arxiu seleccionat per a TA.

El primer model encoder-decoder s'ha definit amb una arquitectura senzilla basada en capes LSTM:

- Capa embedding: S'ha definit amb un vector d'embedding de dimensió 200. Proporcionant un equilibri raonable entre la capacitat de representació semàntica i la càrrega computacional.

- Unitats LSTM: S'han fet servir 100 unitats, buscant un compromís entre la capacitat de càlcul, temps d'entrenament i complexitat del model. A la vista dels resultats obtinguts, potser era necessari triar valors més alts (entre 256 i 512), ja que a l'obtindre un model més complex, es podrien haver captat millor les dependències a llarg termini i les relacions semàntiques complexes.

- Longitud màxima de seqüència: Inicialment s'ha iniciat a 8, però aquest és un valor molt baix. Un valor més òptim seria al voltant del percentil 95 de la longitud de frases del joc de dades.

2. En el primer model de TA, descriu quin efecte tenen els diferents valors en el rendiment del model. Com podríem millorar els resultats d'aquesta tasca?

- Embedding size: Un embedding més gran captura millor matisos semàntics però incrementa la necessitat de memòria. Augmentar aquesta mida (300 o més) podria aportar millors resultats si hi ha prou dades i recursos computacionals.

- Unitats LSTM: Tot i que no s'han modificat durant la pràctica, s'ha vist que amb 100 unitats el model queda limitat en la capacitat de representació. Això implica una reducció significativa en la qualitat de les traduccions generades (de fet, no s'han aconseguit bons resultats). Incrementar aquest valor ajudaria al model a entendre i reproduir millor les relacions entre paraules.

- Epochs: Tot i que s'han entrenat suficients èpoques, el problema està en la capacitat limitada del model i possiblement en l'optimització limitada per l'entrenament des de zero sense embeddings preentrenats.

- Batch size: Aquest valor afecta en l'estabilitat i velocitat de convergència del model. Un batch més gran pot accelerar l'entrenament i estabilitzar les actualitzacions del gradient, però en alguns casos pot provocar una convergència més lenta o dificultat l'exploració de punts locals. Un valor més petit ofereix actualitzacions més freqüents i un entrenament més dinàmic, tot i que pot implicar més soroll en els càlculs del gradient.

Les millores que jo proposaria serien un augment del nombre d'unitats LSTM, l'ús de embeddings preentrenats (com s'ha fet a continuació) i ajustar el bathc size experimentalment per optimitzar el temps d'entrenament i la convergència del model.

Marc Bracons Cucó

3. Compara els dos models de TA entrenats: Explica quines son les principals diferències entre els dos models entrenats. Quina proporciona millors resultats? Per què?

Hi ha diferències en els models, tant en l'arquitectura com en els resultats.

El primer model fa servir embeddings inicialitzats aleatòriament, cosa que limita considerablament la qualitat de les representacions, ja que el model necessita moltes iteracions per aprendre una representació significativa de cada paraula. A més, amb una quantitat relativament petita d'unitats, aquest model té una capacitat molt baixa per capturar dependències llargues i complexes, que són molt importants en traducció automàtica.

El segon model, fa servir embeddings preentrenats amb GloVe, que ja contentn informació semàntica rellevant extreta d'un gran corpus. Aquesta estratègia proporciona una base molt més sòlida, permetent que el model pugui generalitzar millor amb menys èpoques d'entrenament. Com que els embeddings preentrenats estan congelats, el model només ha d'entrenar les capes LSTM i Dense, reduint així la càrrega computacional. Tot i això els resultats tampoc han sigut bons, segurament a causa de la baixa complexitat de la xarxa.

En resum, l'ús d'embeddgins preentrenats és una tècnica recomanable i eficient per millorar els resultats de tasques de traducció automàtica. Aquesta estratègia facilita una convergència més ràpida i millors traduccions sense necessitats de tindre un increment en la complexitat de computació del model.

3. Detecció de NER y NEL (25%)

A partir de les tasques realitzades a l'apartat 2 de la pràctica, titulada "Detecció de NER i NEL", contesta les preguntes següents:

1. Quins resultats s'obtenen en general en la detecció d'entitats? De què depenen, i quin efecte té el corpus utilitzat en la distribució de tipus d'entitats i en la seva avaluació?

A la secció 2, els resultats obtinguts han sigut en general satisfactoris, obtenint un F-score global aproximadament del 77%. Tot i que hi ha variabilitat en funció de la categoria específica de les entitats analitzades. Les categories més freqüents o més ben representades dins del corpus d'entrenament sembla que han obtingut resultats superiors.

Podem dir doncs, que els resultats depenen en gran mesura del corpus d'entrenament que fem servir. Així doncs, una representació equilibrada de diferents categories en aquest corpus és de gran importància per a identificar correctament les entitats en contextos diversos. La qualitat i mida del corpus també han estat determinants en l'eficiència i precisió assolides durant l'entrenament del model.

Marc Bracons Cucó

2. Quin tipus d'entitats podem enllaçar amb Wikidata? Què ocorre si tenim una entitat amb un nom ambigu? Què podríem fer per intentar trobar l'enllaç correcte donat cada context?

Pel que fa a l'enllaç d'entitats anomenades, s'ha observat que les entitats reconegudes automàticament es poden enllaçar amb entrades específiques dins d'aquesta base de coneixement, sobretot aquelles referents a persones i organitzacions i llocs. Tot i això, també s'ha vist que poden haver-hi problemes quan una mateixa denominació pot referir-se a diverses organitzacions, com en el cas de «UAB», que en comptes de la Universitat Autònoma de Barcelona, ho ha entès com American University of Beirut. En aquest cas concret es podrien haver implementat tècniques per utilitzar el context i així identificar l'entitat adequada.