Facultat d'Informàtica de Barcelona

GIA UPC

Processament del llenguatge humà

Pràctica 4

Alumnes :
Casanovas Poirier, Anna
Pumares Benaiges, Irene

Tutors :
Turmo Borrás, Jordi
Medina Herrera, Salvador

June 5, 2024

Contents

1 Introducció

L'objectiu d'aquesta pràctica és l'entrenament i avaluació de models de Word-Embeddings i consta de tres parts principals: en primer lloc, l'entrenament de models de Word2Vec amb diferents mides, utilitzant el Catalan General Crawling com a corpus; en segon lloc, l'entrenament d'un model de Similitud de Text Semàntic amb diferents mètodes d'incrustació de paraules; per últim, l'entrenament del mateix model amb embeddings entrenables.

En la primera part, s'analitzarà la influència de la mida del conjunt de dades tot comparant els diferents models obtinguts amb diverses tècniques.

En la segona i tercera part, s'avaluara la capacitat dels diferents models d'incrustació de paraules utilitzant tècniques com One-Hot encoding, Word2Vec pre-entrenat i altres models basats spaCy. Així com la influència de la inicialització dels embedding entrenables en els resultats obtinguts.

Finalment, s'analitzaran els resultats obtinguts amb l'objectiu de comprendre millor l'eficàcia de les diferents tècniques d'incrustació de paraules i la seva aplicabilitat en tasques de processament de llenguatge natural.

2 Models de Word2Vec

En aquesta secció, s'han entrenat quatre models de Word2Vec per a poder treballar amb ells posteriorment. Per a fer-ho, s'ha dividit el corpus 'catalan_general_crawling' en diferents mides (100MB, 500MB, 1GB i complet).

Word2Vec és un algorisme d'aprenentatge profund que utilitza una xarxa neuronal de dues capes per aprendre associacions de paraules a partir d'un corpus de text donat. Com bé indica el seu nom, Word2Vec representa cada paraula única amb un vector. Els vectors es creen mitjançant una funció matemàtica que indica el nivell de similitud semàntica entre les paraules que representen. És a dir, l'algorisme pren com a entrada un corpus de text i la seva sortida és un conjunt de vectors de característiques que representen les paraules d'aquest corpus.

Un cop s'han creat els quatre models, s'ha d'escollir quin d'ells és el més adequat. Per fer-ho, s'han visualitzat diverses paraules utilitzant l'algorisme t-SNE però no s'han observat diferències significatives entre els diferents models, per la qula cosa no s'ha pogut arribar a cap conclusió ferma. Tot i així, s'ha formulat la hipòtesi de que el millor model és el tercer (1GB), ja que si s'analitza el gràfic resultant, les paraules més similars es troben més properes entre si.

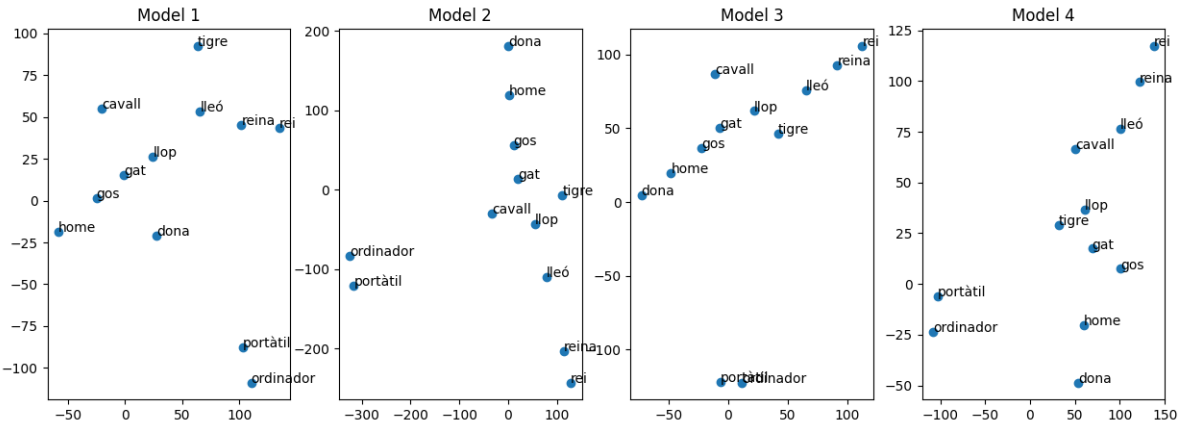

Figure 1: Visualització t-SNE

Per a confirmar la hipòtesis, s'ha decidit realitzar operacions aritmètiques amb els vectors de paraules. Aquesta tècnica permet avaluar la capacitat dels models per capturar relacions semàntiques i analògiques.

S'ha utilitzat el mètode `xv.most_similar` de Word2Vec per a poder fer diverses operacions de suma i resta entre diferents termes. Després de provar diverses operacions, s'ha observat que el model que obté resultats més precisos i coherents en termes generals és el tercer model, reforçant així la hipòtesi inicial.

A més, s'ha usat el mètode `wv.similar_by_word` per trobar les paraules més similars a una paraula donada. En aquest cas, la majoria dels models han funcionat amb èxit, però

el tercer model també ha destacat lleugerament sobre la resta.

Per tant, després d'analitzar els resultats, es pot concloure que el tercer model de Word2Vec, entrenat amb un corpus d'1GB, és el que presenta un millor rendiment general. En conseqüència, aquest model serà l'escollit per utilitzar-lo en les següents seccions d'aquest treball.

3 Model de Similitud de Text Semàntic

A continuació, es faran diferents tipus de word embeddings per després entrenar un model de similitud de text semàntic. Tractarem amb les dades TEXT SIMILARITY de Hugging Face, que compta amb files que tenen dues frases en català i la puntuació de similitud entre elles. S'entrenarà un model bàsic que utilitzarà aquestes representacions vectorials per predir la similitud semàntica entre parells de frases.

Els embeddings de paraules són representacions vectorials de paraules o frases que capturen informació semàntica i sintàctica, i són fonamentals en diverses tasques de Processament del Llenguatge Natural (PLN).

En els següents subapartats, s'exploraran diverses tècniques d'embeddings per comparar la seva efectivitat en la tasca de similitud de text semàntic.

3.1 One-Hot

Aquest mètode representa cada paraula com un vector binari de dimensió igual a la mida del vocabulari, on cada paraula es representa com un vector amb un únic element a 1 (indicant la presència de la paraula) i la resta a 0.

3.2 Models de Word2Vec pre-entrenats

Els models de Word2Vec pre-entrenats són models que s'han entrenat prèviament en l'apartat anterior. Aquests models capten les relacions semàntiques i sintàctiques entre les paraules del corpus, i les representen en forma de vectors densos. Com bé s'ha explciat anteriorment, s'ha escollit el tercer model de Word2Vec entrenat amb un corpus de 1GB com el més adequat.

S'han explorat dos enfocaments per construir el embeddings a partir del model de Word2Vec pre-entrenat seleccionat.

3.2.1 Word2Vec + Mean

Aquest model calcula la representació vectorial d'una frase com la mitjana aritmètica dels vectors de les paraules que la componen. És un enfocament prou bo, tot i que no té en compte la importància relativa de les diferents paraules dins de la frase.

3.2.2 Word2Vec + Mean ponderada (TF-IDF)

En aquest model, la representació vectorial d'una frase es calcula com la mitjana ponderada dels vectors de paraules, on els pesos es deriven de les puntuacions TF-IDF de les paraules, de manera que s'assigna més pes a les paraules menys freqüents, la qual cosa hauria de millorar el rendiment respecte l'anterior.

3.3 spaCy

SpaCy proporciona una eina eficient per processar textos i obtenir embeddings robustos basats en models preentrenats. Els embeddings de SpaCy es generen a partir de models

de llenguatge profund que han estat entrenats en grans corpus de dades de text.

3.4 RoBERTa

RoBERTa és un model transformer avançat que ha estat preentrenat amb grans quantitats de dades textuals per capturar relacions semàntiques profundes. S'ha utilitzat la implementació de RoBERTa en català (roberta-base-ca-v2) a través de SpaCy (ca_core_news_trf). S'han aplicat dues variants explicades a continuació per obtenir embeddings.

3.4.1 CLS

Aquesta tècnica utilitza la representació del token [CLS], que es troba al principi de cada seqüència d'entrada en models basats en transformers, per obtenir un embedding que representa l'oració sencera.

3.4.2 Mean

Aquesta tècnica calcula la mitjana de les representacions vectorials de tots els tokens d'una oració per obtenir un embedding agregat.

3.5 RoBERTa fine-tuned

Per millorar l'especificitat dels embeddings en la tasca de similitud de frases, també s'ha utilitzat un model RoBERTa ajustat específicament per a la tasca d'avaluació semàntica de frases (roberta-base-ca-v2-cased-sts).

4 Model de Similitud de Text Semàntic amb embeddings entrenables

4.1 Random Embeddings

En aquest apartat, s'ha entrenat un model utilitzant embeddings inicialitzats aleatòriament, que es generen mitjançant una distribució uniforme i es van actualitzant durant l'entrenament. Aquest punt de vista parteix de vectors d'incrustació sense cap coneixement preexistent, però permet al model aprendre representacions vectorials específiques per a les dades d'entrenament, ajustant els embeddings per captar millor les relacions semàntiques entre les paraules i les frases del corpus.

4.2 Word2Vec

En aquest apartat, s'utilitzaran embeddings preentrenats amb el model Word2Vec per inicialitzar les representacions vectorials. Aquests embeddings, es prenen com a punt de partida, però el model els anirà actualitzant durant l'entrenament per adaptar-se millor a la tasca específica de similitud de textos.

5 Resultats

S'han executat tots els mètodes descrits a l'apartat anterior. A continuació, s'analitzaran els resultats de tots els models entrenats per a discernir quin d'ells és el millor per a la tasca de similitud de textos.

5.1 Resultats embeddings no entrenables

A continuació es mostra una taula amb els resultats per a cada partició del dataset (train, validation i test) dels models sense els embeddings entrenables.

	PearsonTrain	PearsonVal	PearsonTest
ONE HOT	0.582279	0.401946	0.487937
WORD2VEC - MEAN	0.468767	0.349670	0.412241
WORD2VEC - MEAN PONDER	0.518651	0.369870	0.405008
SPACY	0.522912	0.312013	0.430223
ROBERTA CLS	0.625978	0.183531	0.335894
ROBERTA MEAN	0.732696	0.369659	0.513428
ROBERTA FINE-TUNED	0.781988	0.9474290	0.7522604

Table 1: Resultats dels models de similitud

El model One Hot utilitza representacions binàries de les paraules. Els resultats per al conjunt d'entrenament superen el 0.5, mentres que la validació i el test es mantenen en l'interval 0.4-0.5. Malgrat les limitacions del model per capturar la semàntica, encara és capaç de proporcionar uns resultats millors que altres models en la predicció de la similitud entre frases. No obstant, els seus resultats són inferiors als models més avançats com RoBERTa fine-tuned.

A continuació, el model Word2Vec - Mean utilitza embeddings preentrenats amb el model Word2Vec per representar les paraules. En aquest cas, es fa la mitjana dels vectors d'embeddings de totes les paraules en una frase per obtenir una representació vectorial de la frase. Els resultats es mostren inferiors a la resta de models, el que ens indica que, encara que el Word2vec és capaç de captar informació semàntica de les paraules, la mitjana dels vectors no és suficient per obtenir una representació òptima de la frase i això fa que sigui el pitjor model de tots.

El model Word2Vec - Mean Ponder utilitza una tècnica similar a l'anterior, però en aquest cas, les paraules es ponderen en funció de la seva rellevància. Aquesta tècnica millora lleugerament els resultats respecte a la mitjana simple de Word2Vec, demostrant que aquesta ponderació pot ajudar a capturar millor la informació rellevant de les frases. Tanmateix, encara queda per sota dels resultats obtinguts amb altres models.

El model SpaCy utilitza embeddings preentrenats i també captura relacions semàntiques entre les paraules. Els resultats mostren que el seu rendiment és més consistent que el de Word2Vec, però les diferències en les mètriques indiquen que potser no és tan robust com

altres models preentrenats i afinats específicament per a tasques de similitud de text.

Pel que fa al RoBERTa cls, és un mètode que utilitza el vector de classe CLS generat pel model RoBERTa per representar oracions. Veiem que sobretot al Validation aquest té una correlació de Pearson molt baixa. Això ens pot indicar que el model és tan simple que no és capaç de capturar tota la informació semàntica de les frases. No obstant això, després per la partició del test augmenta una mica tot i que no molt fins a tenir una correlació del 0.33.

Després també s'ha fet un altre mètode utilitzant el model de RoBERTa però fent la mitjana. Aquí veiem una millora significativa respecte l'anterior, ja que el conjunt de test passa del 0.5. Al fer la mitjana dels vectors d'incrustació que es generen millora molt el seu rendiment.

Finalment, tenim el model ROBERTA FINE-TUNED, un model que utilitza el model RoBERTa pre-entrenat que s'ha ajustat específicament per a tasques de similitud de text. Els resultats d'aquest model mostren la millor correlació de Pearson amb diferència. Això ens vol dir que clarament el model s'ha ajustat a la seva fi i ha pogut millorar significativament el seu rendiment en tasques específiques.

5.2 Resultats embeddings entrenables

Un cop s'ha vist l'eficiència dels diferents embedding no entrenables, veurem els resultats de l'últim apartat per poder concluir si ajuda i/o millora el model entrenant els embeddings mentres s'entrena el model. A la taula següent s'observen els resultats:

	PearsonTrain	PearsonVal	PearsonTest
RANDOM (mean)	0.981109	0.200650	0.193280
WORD2VEC	0.951268	0.324359	0.384743
WORD2VEC - cosinus	0.532262	0.330576	0.457132

Table 2: Resultats dels models de similitud

Cal destacar que per als embeddings aleatòris, s'han realitzat tres execucions ja que el caràcter random del model fa que cada cop ens donin valors diferents.

L'anàlisi dels resultats obtinguts amb els embeddings aleatoris (Random Embeddings) i els embeddings pre-entrenats amb Word2Vec ofereix una visió reveladora sobre l'efecte de la inicialització dels vectors d'incrustació en un model de similitud de text semàntic.

Amb els Random Embeddings (uniforme), observem una alta correlació de Pearson en el conjunt d'entrenament, que oscil·la al voltant de 0.978, indicant una bona capacitat del model per ajustar-se a les dades d'entrenament. No obstant això, en el conjunt de validació i de prova, la correlació és significativament més baixa, que no arriba al 0.2 en el cas del test. Aquesta discrepància pot suggerir un problema d'adaptació excessiva (overfitting), ja que el model pot estar capturant massa els detalls del conjunt d'entrenament i no generalitzant adequadament. El fet d'entrenar els embeddings alhora de l'entrenament del model de similitud efectivament ens està fent tenir un overfitting més gran que en

qualsevol dels altes incrustaments de paraules.

D'altra banda, en el cas de l'anàlisi utilitzant o no la distància del cosinus, s'observa una diferència significativa en el rendiment del model entrenat amb Word2Vec. Sense utilitzar la distància del cosinus, el model mostra una correlació de Pearson més alta en el conjunt d'entrenament (0.951) en comparació amb el model que utilitza la distància del cosinus (0.532). Aquesta diferència suggereix que l'ús de la distància del cosinus pot conduir a un sobreajustament a les dades d'entrenament, mentre que sense utilitzar-la, el model generalitza millor a noves dades. Malgrat això, en els conjunts de validació i prova, el model que utilitza la distància del cosinus mostra un rendiment lleugerament millor, amb correlacions de Pearson de 0.331 i 0.457 respectivament, comparat amb el model sense utilitzar-la.

En resum, aquesta anàlisi ens indica que els Word2Vec Embeddings que hem entrenat nosaltres, tot i tenir overfitting, ofereixen una millor generalització i una representació semàntica més rica que els embedding random. A més l'ús de la distància del cosinus fa que tingui un rendiment inclús una mica millor.

Addicionalment, també es pot fer una comparació entre el models Word2Vec que hem entrenat en el apartat anterior i aquests. En el cas dels anteriors arribàven a una correlació de Pearson d'un 0.4 i 0.41 respectivament (per tant no són dels millors models que hem vist). En canvi, utilitzant els embeddings entrenables s'ha arribat a una correlació lleugerament més alta (del 0.45). Tot i això, podem concluir que el pitjor de tots ha estat l'entrenable sense el cosinus ja que ha patit de manera molt extrema d'overfitting, i ha acabat sent pitjor que la resta de models.

6 Conclusions

Després d'analitzar els resultats i comparar els diferents models d'incrustació de paraules en la tasca de similitud de text semàntic, podem arribar a diverses conclusions.

En quant a l'eficàcia dels embeddings no entrenables, podem dir que els que millor rendiment donen són els proporcionats per RoBERTa. Aquesta millora suggereix que els embeddings pre-entrenats contenen informació semàntica més rica i generalitzable, que és fonamental per tasques de similitud de text semàntic. Tot i això, ha sorprès el relativament bon rendiment que ha tingut el one hot encoding, ja que s'esperàven resultats molt més febles per la seva part.

Ha estat obvi que el fine-tuned model amb RoBERTa ha tingut un rendiment extremadament millor als altres, i això ens recorda a la importància de l'ajustament específic de tasques, ja que ens permet adaptar els nostres embedding a les necessitats particulars de la tasca en qüestió.

Després s'ha passat a fer l'entrenament dels embeddings durant el model. L'entrenament simultani dels embeddings durant l'entrenament del model de similitud de text semàntic sembla conduir a un sobreajustament significatiu, com es veu en els resultats amb embeddings aleatoris i els embeddings de Word2Vec. Això suggereix que, en general, és més eficaç utilitzar embeddings pre-entrenats i ajustar-los específicament per a la tasca, en lloc d'entrenar els embeddings durant el procés d'entrenament del model.

Un altre punt important del treball ha estat observar l'impacte de les tècniques d'agregació de vectors. S'han utilitzat tècniques com la mitjana, la mitjana ponderada (TF-IDF) per millorar el rendiment dels nostres embeddings preentrenats.

En resum, aquest treball demostra la importància de seleccionar i ajustar els embeddings de paraules de manera adequada en tasques de PLN, com la similitud de text semàntic, per obtenir resultats òptims. Els embeddings pre-entrenats, especialment quan s'ajusten específicament per a la tasca en qüestió, mostren un rendiment superior en comparació amb altres mètodes d'incrustació de paraules. A més, és crucial evitar el sobreajustament en el procés d'entrenament utilitzant tècniques adequades d'agregació de vectors i ajustament específic de tasques.