WORD EMBEDDINGS

Práctica 4 - Processment del Llenguatge Humà

Lola Monroy Mir i Marta Nadal Par

5 de Juny de 2024

Universitat Politècnica de Catalunya

Facultat d'Informàtica de Barcelona

Índex

1 Introducció

En aquesta pràctica, es realitza la comparació de diferents models i tècniques d'embeddings de paraules, dissenyats per mesurar la similitud semàntica entre textos. Desde els models Word2Vec, tant a la seva versió de mitjana simple com a la ponderada per TF-IDF, fins als models més avançats com RoBERTa en les seves variants CLS i Mean, són analitzats i comparats en termes de la seva capacitat per capturar les relacions semàntiques. Aquests models s'implementen i s'avaluen utilitzant un corpus de dades extens, amb l'objectiu d'entendre com l'elecció de la tècnica d'embeddings i la mida del model afecten la precisió i la robustesa de les representacions semàntiques.

Cal notar que aquest document recull la teoria i el comentari dels resultats, però el desenvolupament del codi que implementa els models i els respectius comentaris sobre aquest, es troben en el notebook que forma part de l'entrega.

2 Entrenar models de Word2Vec

En aquest apartat s'entrenen tres models Word2Vec amb diferents mides del dataset (100MB, 500MB i 1GB), amb l'objectiu de comparar, en l'apartat següent, com la mida dels datasets influeix en el model de Similitud de Text Semàntic. Per fer-ho, s'utilitza el corpus Catalan General Crawling.

Word2Vec és una de les implementacions d'incrustació de paraules més populars. Aquest mètode crea representacions de les paraules en forma de vectors numèrics, capturant tant les qualitats semàntiques com les sintàctiques de les paraules.

Inicialment, es va intentar entrenar el model Word2Vec amb el dataset complet. No obstant això, aquesta tasca va resultar en un error de memòria. Per tant, l'entrenament s'ha limitat a les mides de 100MB, 500MB i 1GB.

Corpus Catalan General Crawling

El Corpus Català General Crawling és un conjunt de dades en català utilitzat per crear models de processament del llenguatge natural. Aquest corpus s'ha obtingut mitjançant la tècnica de "web crawling", que implica la recopilació automàtica de contingut textual de diferents fonts web (notícies, blocs, fòrums, llocs governamentals...). Això garanteix que el corpus capturi exemples del llenguatge en diferents contextos i registres. A més a més, conté una gran quantitat de dades, el que és útil per entrenar models de NLP robusts.

3 Entrenar models de Similitud de Text Semàntic

L'anàlisi de sentiments és una tasca de processament del llenguatge natural que detecta emocions i opinions expressades en un text. Les seves aplicacions van des de l'anàlisi de ressenyes al sentiment de les xarxes socials. En aquesta secció explorarem i compararem quatre tècniques de preprocessament

de text: One-Hot Encoding, Word2Vec, SpaCy i RoBERTa.

Per tal de comparar els resultats dels diferents models de similitud de text semàntic d'incrustació de paraules que hem entrenat, hem utilitzat una estructura base que es repeteix en tots.

3.1 One-Hot

En el context de l'anàlisi de sentiments, on busquem determinar el sentiment o la polaritat emocional de les dades del text, podem representar paraules i frases com a vectors numèrics. Assignem un índex únic a cada categoria paraula o frase, creant un vector. El problema d'aquest mètode, és que no captura similituds semàntiques entre paraules. Totes les paraules són tractades com a completament diferents i independents entre si; aquesta falta de context es veurà reflectida en les dades.

	Corr. Pearson
Train	0.934
Validation	0.142
Test	0.193

La correlació de Pearson a les dades de l'entrenament és extremadament elevada (0.934); hi ha una correlació molt alta entre les prediccions del model i els valors reals. Això, si anés de la mà amb les correlacions a les particions de validació i té prova, seria molt bon senyal, però com que les correlacions de Pearson per aquests últims són molt més baixes (ambdues per sota del 0.2), és indicador d'overfitting. El model està memoritzant les dades d'entrenament en comptes de buscar patrons generalitzables.

Problema en la captura de relacions semàntiques

A més a més, One-hot encoding no captura similituds semàntiques entre paraules. Totes les paraules són tractades com a completament diferents i independents entre si. Com que en cada vector només un element és 1 (indicant la presència de la categoria) i tots els altres són 0, no hi ha espai per la captura de les relacions entre paraules, que és crucial a l'hora de predir el context. Això podria explicar la dificultat de generalitzar del model, i els baixos coeficients de correlació de Pearson pels conjunts amb els quals no ha estat entrenat.

Problema de la dimensionalitat

Per a vocabularis grans, la representació one-hot es torna inpràctica a causa de l'alta dimensionalitat i la dispersió dels vectors. Per això, l'entrenament del model ha trigat hores. En models més simples amb mida limitada, pot ser seria prou eficient, però en aquest cas on les dades arriben a ocupar gigues, no és indicat.

3.2 Models Word2Vec

Word2Vec es basa en la idea que les paraules que es produeixen en contextos similars tendeixen a tenir significats similars. Aprèn aquestes relacions contextuals entrenant una xarxa neuronal en un gran

corpus de dades de text. L'algorisme té en compte les paraules que envolten una paraula objectiu i intenta predir la probabilitat d'observar aquestes paraules de context donada la paraula objectiu. En aquest procés, ajusta els paràmetres interns de la xarxa neuronal per optimitzar les seves prediccions.

L'anàlisi semàntica utilitzant Word2Vec implica mesurar la similitud entre paraules calculant la similitud entre les seves incrustacions de paraules. Les paraules semblants tindran una similitud de cosinus més alta o una distància euclidiana més petita. Això permet fer tasques com ara l'analogia de paraules i l'agrupació de paraules. Per exemple, restant el vector de "rei" de "home" i afegint el vector de "dona", és probable que el vector resultant estigui a prop del vector de "reina", demostrant que Word2Vec captura les relacions de gènere.

Per tal d'avaluar els models i comprovar que funcionin correctament s'ha buscat les paraules més similars a informàtica i s'ha buscat la similitud entre informàtica i coordinador. Observant els resultats obtinguts al 'notebook', veiem que a mesura que la mida del model augmenta, les paraules similars a informàtica tendeixen a tenir una similitud de cosinus més alta, cosa que indica que els models més grans poden capturar millor les relacions semàntiques entre paraules. No obstant això, s'observa una disminució en la similitud entre 'informàtica' i 'coordinador' a mesura que augmenta la mida, el que podria indicar una menor capacitat per capturar relacions específiques entre paraules.

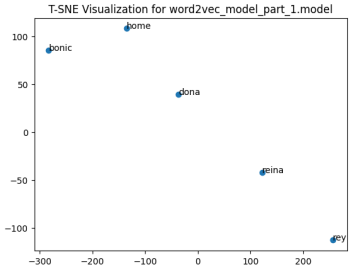

Figure 1: 100MB

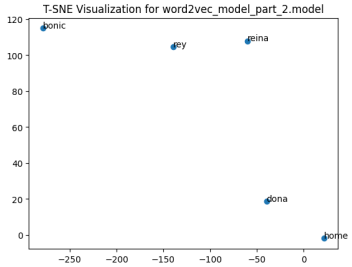

Figure 2: 500MB

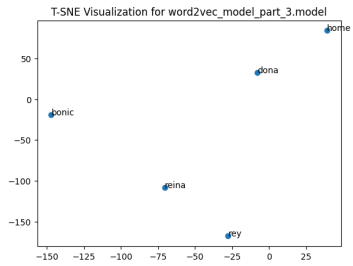

Figure 3: 1GB

A continuació podem observar els resultats obtinguts amb el t-NSE (stochastic Neighbor Embedding), que redueix la dimensionalitat dels embeddings de paraules a dues dimensions, el que permet visualitzar la similitud semàntica entre les paraules.

Observem una consistència en la relació entre rei i reina en els tres models, cosa que suggereix que aquesta relació és robusta i es captura correctament independentment de la mida del model. Però, la variabilitat en la posició d'home, dona i bonic entre les gràfiques indica que la mida del model pot influir en la precisió i consistència d'altres relacions semàntiques.

3.2.1 Word2Vec + Mean

A l'enfocament de Word2Vec + Mean, s'utilitza el model Word2Vec per obtenir representacions vectorials (embeddings) de les paraules en un text. Després, aquests embeddings es fan una mitjana per

obtenir una única representació vectorial del text complet. Això proporciona context a les representacions de les frases.

	Mida del model Word2Vec utilitzat:		
Corr. Pearson	**100 MB**	**500 MB**	**1 GB**
Train	0.611	0.603	0.573
Val	0.485	0.423	0.389
Test	0.541	0.533	0.477

A mesura que augmenta la mida del model, la correlació de Pearson disminueix lleugerament en tots els conjunts. Això pot ser degut al fet que els models més grans capturen més informació, però també poden incloure més soroll o informació irrellevant, afectant la mitjana dels embeddings. Amb això podem inferir que una mida intermèdia pot ser suficient per capturar les relacions semàntiques sense introduir gaire complexitat.

El més important que hem d'extreure d'aquí és que els models més grans poden no necessàriament traduir-se en un millor rendiment, com s'observa als resultats, i poden requerir més recursos computacionals sense beneficis proporcionals.

3.2.2 Word2Vec + Mean ponderada (TF-IDF)

TF-IDF (Term Frequency-Inverse Document Frequency) és una tècnica de representació numèrica que té com a objectiu captar el significat dels termes dins del document en el context de la col·lecció de documents.

Podem segmentar la definició en dues parts. La primera, TF, representa la freqüència amb què apareix una paraula en un document. Es calcula dividint el nombre de vegades que apareix una paraula en un document pel nombre total de paraules del document. En segona instància, IDF vol dir freqüència inversa del document. Aquesta mesura la raresa o la importància d'un terme en tota la col·lecció de documents. Es calcula prenent el logaritme de la relació entre el nombre total de documents i el nombre de documents que contenen el terme.

En combinar els components TF i IDF, TF-IDF assigna pesos més alts a les paraules que són freqüents en un document i poc comuns a tota la col·lecció. Això permet a TF-IDF emfatitzar les paraules que tenen més probabilitats de transmetre informació específica sobre el document.

	Mida del model Word2Vec utilitzat:		
Corr. Pearson	**100 MB**	**500 MB**	**1 GB**
Train	0.616	0.606	0.593
Val	0.395	0.362	0.373
Test	0.469	0.452	0.450

En primer lloc, observem que els valors de la correlació de Pearson varien segons la mida del model Word2Vec utilitzat. Això suggereix que la mida del model té un impacte en el rendiment del model.

Pel conjunt d'entrenament, els valors són relativament alts en totes les mides, amb valors entre 0.593 i 0.616. Això suggereix que el model és capaç de capturar eficaçment la similitud semàntica entre les oracions del conjunt de 'train'.

En el de validació i prova, veiem que disminueixen els valors en comparació amb el conjunt d'entrenament, cosa que ens indica que el model és possible que experimenti una mica de sobreajustament, ja que el seu rendiment és menor en dades que no ha vist durant l'entrenament.

3.3 SpaCy

SpaCy és una biblioteca que proporciona un tokenitzador molt eficient i precís. També disposa d'un lemmatitzador que redueix les paraules a la forma base o lema. Això ens ajuda a normalitzar el text i tractar totes les variants d'una paraula. Aquestes i altres coses, com l'etiquetació PoS o el reconeixement d'entitats anomenades, són atributs de SpaCy.

Per al nostre model, hem obtingut els següents resultats per a cada conjunt.

	Corr. Pearson
Train	0.522
Validation	0.311
Test	0.430

La correlació de Pearson al conjunt d'entrenament és de 0.522, cosa que indica una correlació moderadament bona entre les prediccions del model i els valors reals. El model ha après alguns patrons presents a les dades d'entrenament, però no ha sobre ajustat completament aquestes dades.

Per al conjunt de validació, és de 0.311. És bastant pitjor que l'anterior, per tant, el model no està generalitzant gaire bé dades no vistes durant l'entrenament. Com que el seu rendiment en dades noves no és tan bo com en les dades d'entrenament, pot ser que hi hagi una mica de sobreajustament.

Finalment, la correlació de Pearson al conjunt de prova és de 0.430. Aquest valor és més alt que el de la validació, però encara és menor que el d'entrenament. Amb això, podem concloure que el model té capacitat per generalitzar noves dades, si bé encara hi ha marge de millora. La diferència entre els conjunts de validació i prova també pot indicar variabilitat en les dades o que el model necessita ajustaments addicionals.

3.3.1 Possibles millores per al model Spacy: Eliminar *stopwords*

Spacy construeix *embeddings* de frases fent la mitjana de les incrustacions de paraules. Com que, en una frase hi ha moltes paraules sense sentit (anomenades ***stopwords***), si no s'eliminen poden obtenir-se resultats poc acurats. Per això, és necessari eliminar-les durant el preprocessament.

Entrarem més en profunditat en el punt fet en el paràgraf anterior per acabar-ho d'il·lustrar. Cada paraula té una representació vectorial, apresa mitjançant embeddings contextuals (Word2Vec), que

s'entrenen en els corpus. L'*embedding* de paraules d'una frase completa és simplement la mitjana de totes les paraules diferents. El problema està en què si tenim moltes paraules que es troben semànticament a la mateixa regió (com, per exemple, paraules com "ell", "era", "això", ...) i el vocabulari addicional "s'anul·la", aleshores podria acabar amb una similitud molt més alta de la que realment existeix.

3.4 RoBERTa

RoBERTa vol dir *Robustly Optimised BERT Approach*. Tal com el seu nom indica, aquest és una optimització de BERT (Bidirectional Encoder Representations from Transformers), que malgrat tenir un rendiment excel·lent, es va continuar experimentant amb la seva configuració obtenint mètriques encara millors. Aquest model està entrenat prèviament en una combinació de cinc conjunts de dades massius de 160 GB de dades de text.

Hi ha dues tècniques principals: RoBERTa CLS i RoBERTa Mean. Aquestes difereixen en la manera com extreuen la informació de la seqüència d'entrada processada pel model. Les veurem en les seccions que segueixen.

3.4.1 RoBERTa CLS

Aquesta tècnica és especialment útil per a tasques de classificació, ja que existeix un token [CLS] entrenat per afegir la informació contextual de tota la seqüència. Per exemple, en tasques com la classificació de sentiments o la detecció de correu brossa, la representació obtinguda del token [CLS] pot ser molt efectiva.

	Corr. Pearson
Train	0.782
Validation	0.220
Test	0.365

En aquest cas, veiem que probablement s'hagi produït un sobreajust a les dades d'entrenament, ja que tot i que la correlació per a les dades d'entrenament és molt alta (0.782), quan veu dades noves li costa generalitzar, com veiem a les correlacions obtingudes de la partició de validació i la de prova.

Per solucionar-ho, podríem utilitzar tècniques addicionals per millorar la capacitat de generalització, com poden ser l'ús de regularització (e.g. L2 regularització), l'augment de dades, o l'ajust d'hiperparàmetres.

3.4.2 RoBERTa Mean

RoBERTa Mean, en lloc de dependre d'un únic token, pren la sortida de l'última capa per a tots els tokens a la seqüència i calcula la mitjana (mean) d'aquests embeddings. Això permet capturar millor la informació distribuïda al llarg de tota la seqüència, cosa que pot ser beneficiós en seqüències llargues on la informació rellevant no es concentra en un sol token. Aquesta tècnica és útil en tasques la generació de resums, on tenir en compte tot el text és molt important.

	Corr. Pearson
Train	0.912
Validation	0.408
Test	0.563

Veiem que aquest model mostra un rendiment molt millor que l'anterior. Tot i que també hi ha un sobreajustament notable a les dades d'entrenament (ja que hi ha una correlació de Pearson del 0.912 per al test, que està lluny de les del train i el val), rendeix notablement millor que el RoBERTa CLS. Per al test tenim una correlació del 0,563, que és prou bona.

Per tant, en el que fa als models RoBERTa CLS i Mean, per a la nostra tasca rendeix millor el que treballa amb les mitjanes, probablement perquè el que el model tingui més context, afavoreix la qualitat de les prediccions.

3.5 RoBERTa fine-tuned

Aquest model *fine-tuned*, ha estat preentrenat; RoBERTa s'ajusta per a una tasca específica usant un conjunt de dades etiquetatge. Durant aquest procés, els pesos del model s'optimitzen per millorar el rendiment del model. Per tant, només tenim la partició del test, ja que tant el train com el val només tenen sentit per l'entrenament.

La correlació de Pearson que obtenim per al conjunt de dades de prova és: **0.7819886697756575**

És un resultat extraordinàriament bo pel test, que en aquesta pràctica encara no hem assolit. Per tant, podem dir que el rendiment d'aquest model és remarcablement bo. Fent recerca sobre per què ha funcionat tan bé, hem après que durant el procés de fine-tuning, els pesos del model s'ajusten específicament per a la tasca objectiu utilitzant un conjunt de dades etiquetat. Això permet al model capturant millor els patrons rellevants i optimitzar el seu rendiment.

A més a més, aquest procés de fine-tuning pot incloure tècniques de regularització (que hem mencionat en la secció anterior que podien millorar el rendiment) per millorar la capacitat de generalització del model.

3.6 Conclusions: Models de Similitud de Text Semàntic

En la nostra opinió, la conclusió més important a extreure de la comparació entre els models, és la consideració de la importància de cada paraula en el context del text, o almenys no ignorar el context de la paraula. Com va ilustrar John Rupert Firth amb la frase *you shall know a word by the company it keeps*, el context és fonamental a l'hora de fer prediccions sobre textos. Per això, models que fan servir One-Hot encoding que només consideren la paraula sense les seves relacions, rendeixen molt pitjor que altres que usen RoBERTa, considerant el context complet.

4 Model amb Embeddings Entrenables

En aquest cas, s'entrena els models utilitzant les dades d'entrenament durant 128 èpoques. Les dades de validació, com en casos anteriors s'utilitzen per avaluar el rendiment del model després de cada època.

En aquesta secció de la pràctica, inicialitzarem el model amb embeddings entrenables, utilitzant tant Word2Vec com amb Random Embeddings. El model Word2Vec preentrenat que utilitzarem és el de mida 1GB, ja que, com hem observat en apartats anteriors, proporcina una major consistència semàntica, una millor representació de paraules i una millor generalització de les dades no vistes, reduint així l'overfitting.

Pel que fa als embeddings aleatoris, realitzarem diverses iteracions i calcularem la mitjana dels resultats per assegurar-nos que els resultats obtinguts no siguin producte de l'atzar, sinó que reflecteixin de manera acurada el rendiment del model amb embeddings aleatoris.

	Embeddings entrenables Inicialitzats amb:				
Corr. Pearson	**Word2vec**	Rand. (iteració 1)	Rand. (i2)	Rand. (i3)	**Random** (mitjana)
Train	0.498	0.995	0.995	0.995	0.995
Val	0.656	0.248	0.179	0.351	0.259
Test	0.465	0.458	0.395	0.443	0.432

Els embeddings inicialitzats aleatòriament en diferents iteracions mostren una correlació alta de Pearson (exactament 0.995 en tots els casos), cosa que suggereix un sobreajust a les dades d'entrenament. En canvi, els embeddings inicialitzats amb Word2Vec mostren una correlació menor (0.498), la qual cosa suggereix el contrari, un subajustament. Aquest fet pot ser degut al fet que els embeddings de Word2Vec, encara que preentrenats i útils per capturar relacions semàntiques generals, poden no estar adaptats a aquest conjunt específic de dades.

La correlació de Pearson per a les dades de validació és molt variable per al cas dels embeddings inicialitzats aleatoriament, però sempre més baixa que 0.4, així que sigui com sigui sempre és dolenta. Aquest cas evidencia la necessitat de fer vàries mostres del mateix model, ja que al contrari del cas anterior, els resultats poden ser bastant variants. La correlació baixa en les dades de validació suggereix que el model no generalitza bé quan s'entrena amb embeddings aleatoris. La variabilitat entre les iteracions indiquen que el rendiment és dependent de la inicialització aleatòria.

Els embeddings inicialitzats amb Word2Vec tenen una millor correlació (0.656) comparada amb els inicialitzats aleatòriament, la mitjana dels quals és 0.259. Aquesta correlació més alta a les dades de validació indica que els embeddings de Word2Vec estan generalitzant millor en dades no vistes.

La correlació per al conjunt test amb els embeddings aleatoris, no és massa bona, però és millor que la de la validació. De nou, queda reflectida la capacitat limitada per generalitzar dades noves.

Finalment, pel que fa al conjunt test per a la inicialització amb Word2Vc, la correlació no és especialment baixa, per tant, tot i que podria ser millorada, els embeddings de Word2Vec també generalitzen prou bé a dades completament noves, encara que no tan bé com en les dades de validació.

4.1 Conclusions: Models amb embeddings entrenables

En primer lloc, hem conclòs que els embeddings preentrenats amb Word2Vec mostren una millor capacitat de generalització en dades de validació i prova, tot i que no hi ha una diferència tan gran entre aquells models inicialitzats amb Word2Vec o aleatòriament. Això pot ser perquè Word2Vec captura relacions semàntiques generals durant el seu preentrenament, la qual cosa ajuda el model a evitar el sobreajustament i a aprendre representacions útils que són aplicables a noves dades.

En segon lloc, es veu clarament que l'alt coeficient de correlació a les dades d'entrenament per als embeddings aleatoris indica sobreajustament. El model està memoritzant les dades d'entrenament en lloc d'aprendre patrons generals, cosa que provoca un mal rendiment en les dades de validació i prova.

A més a més, una conclusió valuosa a què hem arribat, és que la variabilitat dels coeficients de correlació mostra que aquests resultats són altament dependents de la inicialització. Per tant, els embeddings aleatoris no són una bona elecció per a aquest problema.

L'ús d'embeddings preentrenats com Word2Vec proporciona una base sòlida per a la generalització a noves dades, mentre que els embeddings aleatoris poden portar a problemes de sobreajustament i resultats inconsistents.

5 Bibliografia

1. Comparing Text Preprocessing Techniques: One-Hot Encoding, Bag of Words, TF-IDF, and Word2Vec for Sentiment Analysis. (s/f). https://medium.com/@rayanimran307/comparing-text-preprocessing

2. Presentació numero 11. https://gebakx.github.io/plh/s9/index.html#2

3. What is Word2vec? (s/f). *Mathworks.com*. https://es.mathworks.com/discovery/word2vec.html

4. SpacyStrange similarity between two sentences. (s/f). *stackoverflow.com*. Recuperado de https://stackoverflow.com/questions/52113939/spacy-strange-similarity-between-two-sentences

5. Catalan General Crawling. *Huggingface.co*. https://huggingface.co/datasets/projecte-aina/catalan_general_crawling

6. Corrector ortogràfic i gramatical en català. *Softcatalà*.https://www.softcatala.org/corrector/

7. Overleaf. *Overleaf*. https://www.overleaf.com/