

# Database-driven Chatbot

Eduardo Neves  
Universidade de Coimbra  
email@example.com

## Abstract

No panorama atual vemos em várias plataformas algoritmos de comunicação com pessoas. Entre muitas aplicações, estes podem ser encontrados em vários *websites* banais. A sua utilidade no mundo empresarial viu até crescer mecanismos de construção personalizada de *Chatbots*. Neste trabalho planeia-se desenvolver os mecanismos para a construção de um *Chatbot AI* a partir de *Machine Learning*.

## 1 Introdução

A Chatbot is

## 2 Dados e Abordagem

Os dados para um *Chatbot* baseiam-se principalmente em sequências de interações entre partes. Os *datasets* podem variar entre pergunta e resposta, e e-mails trocados ou conversas em plataformas online entre indivíduos. Para este trabalho foi usado um conjunto de dados baseado em legendas de filmes para português.

### 2.1 Conjunto de dados

O pacote de dados usado foi tirado do projeto "OPUS ... the open parallel corpus" [OPUS, 2022], um 'corpus' de textos traduzidos da internet baseada em produtos 'open source'. Este conjunto é atualmente distribuído pela plataforma "opensubtitles" [opensubtitles, 2022], de onde foi retirado o conteúdo.

Como o foco do projeto é um algoritmo em português, retirou-se o dataset "pt", que coleciona mais de 500000 ficheiros de legendas de filmes até o ano de 2017. Separados por anos, alguns filmes contam com várias versões do mesmo filme. O seu tratamento é explicado na secção seguinte.

### 2.2 Tratamento dos dados

Com o conjunto de dados inicial com redundâncias e pouco estruturado, procurou-se correr um pequena organização à informação. A organização do fluxo de entrada para treino é feita aquando a inicialização do processo.

```
<?xml version="1.0" encoding="UTF-8" ?>
<time value="200011112200" id="14.0" />
<u xpos="14.1" head="14.2" feats="Gender-Masc|Number-Sing|PronType-Ind" upos="DET" lemma="outro" id="14.1" deprel="det">outro</u>
<u xpos="14.2" head="14.9" feats="Gender-Masc|Number-Sing" upos="NOUN" lemma="dia" id="14.2" deprel="nsubj">dia</u>
<u xpos="14.3" head="14.4" upos="CCONJ" lemma="e" id="14.3" deprel="cc">e</u>
<u xpos="14.4" head="14.2" feats="Gender-Fem|Number-Sing" upos="PROPN" lemma="Brecca" id="14.4" deprel="con">Brecca</u>
<u xpos="14.5" head="14.9" feats="Noun-Ind|Number-Sing|Person-3|Tense-Pres|VerbForm-Inf" upos="AUX" lemma="está" id="14.5" deprel="cop">está</u>
<u xpos="14.6" head="14.9" feats="Gender-Fem|Number-Sing|PronType-Ind" upos="NOUN" lemma="cada" id="14.6" deprel="mod:mod">cada</u>
<u xpos="14.7" head="14.8" feats="Gender-Fem|Number-Sing" upos="ADP" lemma="vez" id="14.7" deprel="case">vez</u>
<u xpos="14.8" head="14.6" upos="NOUN" lemma="mais" id="14.8" deprel="mod">mais</u>
<u xpos="14.9" head="14.9" upos="NOUN" lemma="porto" id="14.9" deprel="root">porto</u>
<u xpos="14.10" head="14.9" upos="ADP" lemma="de" id="14.10" deprel="case">de</u>
<u xpos="14.11" head="14.12" upos="ADP" lemma="de" id="14.11" deprel="case">de</u>
<u xpos="14.12" head="14.12" feats="Definite-Def|Gender-Fem|Number-Sing|PronType-Art" upos="DET" lemma="a" id="14.12" deprel="det">a</u>
<u xpos="14.13" head="14.9" feats="Gender-Fem|Number-Sing" upos="NOUN" lemma="vitoria" id="14.13" deprel="obl">vitoria</u>
<u xpos="14.14" head="14.9" upos="PUNCT" lemma="." id="14.14" deprel="punct">.</u>
tag = time (close)
```

Figure 1: Formato original do *dataset* para um filme aleatório

```
"line 1": "uma série original, BETELIA STRANGER STRANGER não foi eu l foi o", "line 2": "lee", "line 3": "é e que o meu", "line 4": "fim de confiar em", "line 5": "lee", "line 6": "Dis que ele lhe deu um telemóvel inútil e lhe disse para o atirar a água ? \". Este é o verdadeiro ", "line 7": "Coloque os algarves em o local ", "line 8": "Diz que o Sr ", "line 9": "Lee fez isso ? Duve ", "line 10": "Achas que eu queria fazer isto ? Pensa em o que", "line 11": "fazer isto me fez sentir ", "line 12": "Tiguel mais de o que frustrado ", "line 13": "O que é que pode provar que o Sr ", "line 14": "Lee é o responsável ?", "line 15": "Não se vê logo ? Isto é necessário entre nós ? Então, Huang, 55 anos, 1 Vozes ser sinceros um com o outro, de homem para homem ", "line 16": "Até quando temos", "line 17": "de tratar os instintos por as razões e por as provas ? 55 sincero ", "line 18": "Sei que também acho que ele é o responsável ", "line 19": "Não é ? Está a ser", "line 20": "mais de o que ridículo ", "line 21": "Como ? Está a fazer essa alegação sem provas ", "line 22": "É ridículo ", "line 23": "Isto é tudo culpa sua ", "line 24": "Nada de isto teria acontecido se não tivesse dito que Park conhecia a rapariga ! Credo ", "line 25": "Até os meus ouvidos ficaram envergonhados com isso", "line 26": "Pode parar ", "line 27": "Por os vícios, goste de ser amado um viril ", "line 28": "Não ", "line 29": "Não ", "line 30": "Não ", "line 31": "Não ", "line 32": "Não ", "line 33": "Não ", "line 34": "Não ", "line 35": "Não ", "line 36": "Não ", "line 37": "Não ", "line 38": "Não ", "line 39": "Não ", "line 40": "Não ", "line 41": "Não ", "line 42": "Não ", "line 43": "Não ", "line 44": "Não ", "line 45": "Não ", "line 46": "Não ", "line 47": "Não ", "line 48": "Não ", "line 49": "Não ", "line 50": "Não ", "line 51": "Não ", "line 52": "Não ", "line 53": "Não ", "line 54": "Não ", "line 55": "Não ", "line 56": "Não ", "line 57": "Não ", "line 58": "Não ", "line 59": "Não ", "line 60": "Não ", "line 61": "Não ", "line 62": "Não ", "line 63": "Não ", "line 64": "Não ", "line 65": "Não ", "line 66": "Não ", "line 67": "Não ", "line 68": "Não ", "line 69": "Não ", "line 70": "Não ", "line 71": "Não ", "line 72": "Não ", "line 73": "Não ", "line 74": "Não ", "line 75": "Não ", "line 76": "Não ", "line 77": "Não ", "line 78": "Não ", "line 79": "Não ", "line 80": "Não ", "line 81": "Não ", "line 82": "Não ", "line 83": "Não ", "line 84": "Não ", "line 85": "Não ", "line 86": "Não ", "line 87": "Não ", "line 88": "Não ", "line 89": "Não ", "line 90": "Não ", "line 91": "Não ", "line 92": "Não ", "line 93": "Não ", "line 94": "Não ", "line 95": "Não ", "line 96": "Não ", "line 97": "Não ", "line 98": "Não ", "line 99": "Não ", "line 100": "Não ", "line 101": "Não ", "line 102": "Não ", "line 103": "Não ", "line 104": "Não ", "line 105": "Não ", "line 106": "Não ", "line 107": "Não ", "line 108": "Não ", "line 109": "Não ", "line 110": "Não ", "line 111": "Não ", "line 112": "Não ", "line 113": "Não ", "line 114": "Não ", "line 115": "Não ", "line 116": "Não ", "line 117": "Não ", "line 118": "Não ", "line 119": "Não ", "line 120": "Não ", "line 121": "Não ", "line 122": "Não ", "line 123": "Não ", "line 124": "Não ", "line 125": "Não ", "line 126": "Não ", "line 127": "Não ", "line 128": "Não ", "line 129": "Não ", "line 130": "Não ", "line 131": "Não ", "line 132": "Não ", "line 133": "Não ", "line 134": "Não ", "line 135": "Não ", "line 136": "Não ", "line 137": "Não ", "line 138": "Não ", "line 139": "Não ", "line 140": "Não ", "line 141": "Não ", "line 142": "Não ", "line 143": "Não ", "line 144": "Não ", "line 145": "Não ", "line 146": "Não ", "line 147": "Não ", "line 148": "Não ", "line 149": "Não ", "line 150": "Não ", "line 151": "Não ", "line 152": "Não ", "line 153": "Não ", "line 154": "Não ", "line 155": "Não ", "line 156": "Não ", "line 157": "Não ", "line 158": "Não ", "line 159": "Não ", "line 160": "Não ", "line 161": "Não ", "line 162": "Não ", "line 163": "Não ", "line 164": "Não ", "line 165": "Não ", "line 166": "Não ", "line 167": "Não ", "line 168": "Não ", "line 169": "Não ", "line 170": "Não ", "line 171": "Não ", "line 172": "Não ", "line 173": "Não ", "line 174": "Não ", "line 175": "Não ", "line 176": "Não ", "line 177": "Não ", "line 178": "Não ", "line 179": "Não ", "line 180": "Não ", "line 181": "Não ", "line 182": "Não ", "line 183": "Não ", "line 184": "Não ", "line 185": "Não ", "line 186": "Não ", "line 187": "Não ", "line 188": "Não ", "line 189": "Não ", "line 190": "Não ", "line 191": "Não ", "line 192": "Não ", "line 193": "Não ", "line 194": "Não ", "line 195": "Não ", "line 196": "Não ", "line 197": "Não ", "line 198": "Não ", "line 199": "Não ", "line 200": "Não ", "line 201": "Não ", "line 202": "Não ", "line 203": "Não ", "line 204": "Não ", "line 205": "Não ", "line 206": "Não ", "line 207": "Não ", "line 208": "Não ", "line 209": "Não ", "line 210": "Não ", "line 211": "Não ", "line 212": "Não ", "line 213": "Não ", "line 214": "Não ", "line 215": "Não ", "line 216": "Não ", "line 217": "Não ", "line 218": "Não ", "line 219": "Não ", "line 220": "Não ", "line 221": "Não ", "line 222": "Não ", "line 223": "Não ", "line 224": "Não ", "line 225": "Não ", "line 226": "Não ", "line 227": "Não ", "line 228": "Não ", "line 229": "Não ", "line 230": "Não ", "line 231": "Não ", "line 232": "Não ", "line 233": "Não ", "line 234": "Não ", "line 235": "Não ", "line 236": "Não ", "line 237": "Não ", "line 238": "Não ", "line 239": "Não ", "line 240": "Não ", "line 241": "Não ", "line 242": "Não ", "line 243": "Não ", "line 244": "Não ", "line 245": "Não ", "line 246": "Não ", "line 247": "Não ", "line 248": "Não ", "line 249": "Não ", "line 250": "Não ", "line 251": "Não ", "line 252": "Não ", "line 253": "Não ", "line 254": "Não ", "line 255": "Não ", "line 256": "Não ", "line 257": "Não ", "line 258": "Não ", "line 259": "Não ", "line 260": "Não ", "line 261": "Não ", "line 262": "Não ", "line 263": "Não ", "line 264": "Não ", "line 265": "Não ", "line 266": "Não ", "line 267": "Não ", "line 268": "Não ", "line 269": "Não ", "line 270": "Não ", "line 271": "Não ", "line 272": "Não ", "line 273": "Não ", "line 274": "Não ", "line 275": "Não ", "line 276": "Não ", "line 277": "Não ", "line 278": "Não ", "line 279": "Não ", "line 280": "Não ", "line 281": "Não ", "line 282": "Não ", "line 283": "Não ", "line 284": "Não ", "line 285": "Não ", "line 286": "Não ", "line 287": "Não ", "line 288": "Não ", "line 289": "Não ", "line 290": "Não ", "line 291": "Não ", "line 292": "Não ", "line 293": "Não ", "line 294": "Não ", "line 295": "Não ", "line 296": "Não ", "line 297": "Não ", "line 298": "Não ", "line 299": "Não ", "line 300": "Não ", "line 301": "Não ", "line 302": "Não ", "line 303": "Não ", "line 304": "Não ", "line 305": "Não ", "line 306": "Não ", "line 307": "Não ", "line 308": "Não ", "line 309": "Não ", "line 310": "Não ", "line 311": "Não ", "line 312": "Não ", "line 313": "Não ", "line 314": "Não ", "line 315": "Não ", "line 316": "Não ", "line 317": "Não ", "line 318": "Não ", "line 319": "Não ", "line 320": "Não ", "line 321": "Não ", "line 322": "Não ", "line 323": "Não ", "line 324": "Não ", "line 325": "Não ", "line 326": "Não ", "line 327": "Não ", "line 328": "Não ", "line 329": "Não ", "line 330": "Não ", "line 331": "Não ", "line 332": "Não ", "line 333": "Não ", "line 334": "Não ", "line 335": "Não ", "line 336": "Não ", "line 337": "Não ", "line 338": "Não ", "line 339": "Não ", "line 340": "Não ", "line 341": "Não ", "line 342": "Não ", "line 343": "Não ", "line 344": "Não ", "line 345": "Não ", "line 346": "Não ", "line 347": "Não ", "line 348": "Não ", "line 349": "Não ", "line 350": "Não ", "line 351": "Não ", "line 352": "Não ", "line 353": "Não ", "line 354": "Não ", "line 355": "Não ", "line 356": "Não ", "line 357": "Não ", "line 358": "Não ", "line 359": "Não ", "line 360": "Não ", "line 361": "Não ", "line 362": "Não ", "line 363": "Não ", "line 364": "Não ", "line 365": "Não ", "line 366": "Não ", "line 367": "Não ", "line 368": "Não ", "line 369": "Não ", "line 370": "Não ", "line 371": "Não ", "line 372": "Não ", "line 373": "Não ", "line 374": "Não ", "line 375": "Não ", "line 376": "Não ", "line 377": "Não ", "line 378": "Não ", "line 379": "Não ", "line 380": "Não ", "line 381": "Não ", "line 382": "Não ", "line 383": "Não ", "line 384": "Não ", "line 385": "Não ", "line 386": "Não ", "line 387": "Não ", "line 388": "Não ", "line 389": "Não ", "line 390": "Não ", "line 391": "Não ", "line 392": "Não ", "line 393": "Não ", "line 394": "Não ", "line 395": "Não ", "line 396": "Não ", "line 397": "Não ", "line 398": "Não ", "line 399": "Não ", "line 400": "Não ", "line 401": "Não ", "line 402": "Não ", "line 403": "Não ", "line 404": "Não ", "line 405": "Não ", "line 406": "Não ", "line 407": "Não ", "line 408": "Não ", "line 409": "Não ", "line 410": "Não ", "line 411": "Não ", "line 412": "Não ", "line 413": "Não ", "line 414": "Não ", "line 415": "Não ", "line 416": "Não ", "line 417": "Não ", "line 418": "Não ", "line 419": "Não ", "line 420": "Não ", "line 421": "Não ", "line 422": "Não ", "line 423": "Não ", "line 424": "Não ", "line 425": "Não ", "line 426": "Não ", "line 427": "Não ", "line 428": "Não ", "line 429": "Não ", "line 430": "Não ", "line 431": "Não ", "line 432": "Não ", "line 433": "Não ", "line 434": "Não ", "line 435": "Não ", "line 436": "Não ", "line 437": "Não ", "line 438": "Não ", "line 439": "Não ", "line 440": "Não ", "line 441": "Não ", "line 442": "Não ", "line 443": "Não ", "line 444": "Não ", "line 445": "Não ", "line 446": "Não ", "line 447": "Não ", "line 448": "Não ", "line 449": "Não ", "line 450": "Não ", "line 451": "Não ", "line 452": "Não ", "line 453": "Não ", "line 454": "Não ", "line 455": "Não ", "line 456": "Não ", "line 457": "Não ", "line 458": "Não ", "line 459": "Não ", "line 460": "Não ", "line 461": "Não ", "line 462": "Não ", "line 463": "Não ", "line 464": "Não ", "line 465": "Não ", "line 466": "Não ", "line 467": "Não ", "line 468": "Não ", "line 469": "Não ", "line 470": "Não ", "line 471": "Não ", "line 472": "Não ", "line 473": "Não ", "line 474": "Não ", "line 475": "Não ", "line 476": "Não ", "line 477": "Não ", "line 478": "Não ", "line 479": "Não ", "line 480": "Não ", "line 481": "Não ", "line 482": "Não ", "line 483": "Não ", "line 484": "Não ", "line 485": "Não ", "line 486": "Não ", "line 487": "Não ", "line 488": "Não ", "line 489": "Não ", "line 490": "Não ", "line 491": "Não ", "line 492": "Não ", "line 493": "Não ", "line 494": "Não ", "line 495": "Não ", "line 496": "Não ", "line 497": "Não ", "line 498": "Não ", "line 499": "Não ", "line 500": "Não ", "line 501": "Não ", "line 502": "Não ", "line 503": "Não ", "line 504": "Não ", "line 505": "Não ", "line 506": "Não ", "line 507": "Não ", "line 508": "Não ", "line 509": "Não ", "line 510": "Não ", "line 511": "Não ", "line 512": "Não ", "line 513": "Não ", "line 514": "Não ", "line 515": "Não ", "line 516": "Não ", "line 517": "Não ", "line 518": "Não ", "line 519": "Não ", "line 520": "Não ", "line 521": "Não ", "line 522": "Não ", "line 523": "Não ", "line 524": "Não ", "line 525": "Não ", "line 526": "Não ", "line 527": "Não ", "line 528": "Não ", "line 529": "Não ", "line 530": "Não ", "line 531": "Não ", "line 532": "Não ", "line 533": "Não ", "line 534": "Não ", "line 535": "Não ", "line 536": "Não ", "line 537": "Não ", "line 538": "Não ", "line 539": "Não ", "line 540": "Não ", "line 541": "Não ", "line 542": "Não ", "line 543": "Não ", "line 544": "Não ", "line 545": "Não ", "line 546": "Não ", "line 547": "Não ", "line 548": "Não ", "line 549": "Não ", "line 550": "Não ", "line 551": "Não ", "line 552": "Não ", "line 553": "Não ", "line 554": "Não ", "line 555": "Não ", "line 556": "Não ", "line 557": "Não ", "line 558": "Não ", "line 559": "Não ", "line 560": "Não ", "line 561": "Não ", "line 562": "Não ", "line 563": "Não ", "line 564": "Não ", "line 565": "Não ", "line 566": "Não ", "line 567": "Não ", "line 568": "Não ", "line 569": "Não ", "line 570": "Não ", "line 571": "Não ", "line 572": "Não ", "line 573": "Não ", "line 574": "Não ", "line 575": "Não ", "line 576": "Não ", "line 577": "Não ", "line 578": "Não ", "line 579": "Não ", "line 580": "Não ", "line 581": "Não ", "line 582": "Não ", "line 583": "Não ", "line 584": "Não ", "line 585": "Não ", "line 586": "Não ", "line 587": "Não ", "line 588": "Não ", "line 589": "Não ", "line 590": "Não ", "line 591": "Não ", "line 592": "Não ", "line 593": "Não ", "line 594": "Não ", "line 595": "Não ", "line 596": "Não ", "line 597": "Não ", "line 598": "Não ", "line 599": "Não ", "line 600": "Não ", "line 601": "Não ", "line 602": "Não ", "line 603": "Não ", "line 604": "Não ", "line 605": "Não ", "line 606": "Não ", "line 607": "Não ", "line 608": "Não ", "line 609": "Não ", "line 610": "Não ", "line 611": "Não ", "line 612": "Não ", "line 613": "Não ", "line 614": "Não ", "line 615": "Não ", "line 616": "Não ", "line 617": "Não ", "line 618": "Não ", "line 619": "Não ", "line 620": "Não ", "line 621": "Não ", "line 622": "Não ", "line 623": "Não ", "line 624": "Não ", "line 625": "Não ", "line 626": "Não ", "line 627": "Não ", "line 628": "Não ", "line 629": "Não ", "line 630": "Não ", "line 631": "Não ", "line 632": "Não ", "line 633": "Não ", "line 634": "Não ", "line 635": "Não ", "line 636": "Não ", "line 637": "Não ", "line 638": "Não ", "line 639": "Não ", "line 640": "Não ", "line 641": "Não ", "line 642": "Não ", "line 643": "Não ", "line 644": "Não ", "line 645": "Não ", "line 646": "Não ", "line 647": "Não ", "line 648": "Não ", "line 649": "Não ", "line 650": "Não ", "line 651": "Não ", "line 652": "Não ", "line 653": "Não ", "line 654": "Não ", "line 655": "Não ", "line 656": "Não ", "line 657": "Não ", "line 658": "Não ", "line 659": "Não ", "line 660": "Não ", "line 661": "Não ", "line 662": "Não ", "line 663": "Não ", "line 664": "Não ", "line 665": "Não ", "line 666": "Não ", "line 667": "Não ", "line 668": "Não ", "line 669": "Não ", "line 670": "Não ", "line 671": "Não ", "line 672": "Não ", "line 673": "Não ", "line 674": "Não ", "line 675": "Não ", "line 676": "Não ", "line 677": "Não ", "line 678": "Não ", "line 679": "Não ", "line 680": "Não ", "line 681": "Não ", "line 682": "Não ", "line 683": "Não ", "line 684": "Não ", "line 685": "Não ", "line 686": "Não ", "line 687": "Não ", "line 688": "Não ", "line 689": "Não ", "line 690": "Não ", "line 691": "Não ", "line 692": "Não ", "line 693": "Não ", "line 694": "Não ", "line 695": "Não ", "line 696": "Não ", "line 697": "Não ", "line 698": "Não ", "line 699": "Não ", "line 700": "Não ", "line 701": "Não ", "line 702": "Não ", "line 703": "Não ", "line 704": "Não ", "line 705": "Não ", "line 706": "Não ", "line 707": "Não ", "line 708": "Não ", "line 709": "Não ", "line 710": "Não ", "line 711": "Não ", "line 712": "Não ", "line 713": "Não ", "line 714": "Não ", "line 715": "Não ", "line 716": "Não ", "line 717": "Não ", "line 718": "Não ", "line 719": "Não ", "line 720": "Não ", "line 721": "Não ", "line 722": "Não ", "line 723": "Não ", "line 724": "Não ", "line 725": "Não ", "line 726": "Não ", "line 727": "Não ", "line 728": "Não ", "line 729": "Não ", "line 730": "Não ", "line 731": "Não ", "line 732": "Não ", "line 733": "Não ", "line 734": "Não ", "line 735": "Não ", "line 736": "Não ", "line 737": "Não ", "line 738": "Não ", "line 739": "Não ", "line 740": "Não ", "line 741": "Não ", "line 742": "Não ", "line 743": "Não ", "line 744": "Não ", "line 745": "Não ", "line 746": "Não ", "line 747": "Não ", "line 748": "Não ", "line 749": "Não ", "line 750": "Não ", "line 751": "Não ", "line 752": "Não ", "line 753": "Não ", "line 754": "Não ", "line 755": "Não ", "line 756": "Não ", "line 757": "Não ", "line 758": "Não ", "line 759": "Não ", "line 760": "Não ", "line 761": "Não ", "line 762": "Não ", "line 763": "Não ", "line 764": "Não ", "line 765": "Não ", "line 766": "Não ", "line 767": "Não ", "line 768": "Não ", "line 769": "Não ", "line 770": "Não ", "line 771": "Não ", "line 772": "Não ", "line 773": "Não ", "line 774": "Não ", "line 775": "Não ", "line 776": "Não ", "line 777": "Não ", "line 778": "Não ", "line 779": "Não ", "line 780": "Não ", "line 781": "Não ", "line 782": "Não ", "line 783": "Não ", "line 784": "Não ", "line 785": "Não ", "line 786": "Não ", "line 787": "Não ", "line 788": "Não ", "line 789": "Não ", "line 790": "Não ", "line 791": "Não ", "line 792": "Não ", "line 793": "Não ", "line 794": "Não ", "line 795": "Não ", "line 796": "Não ", "line 797": "Não ", "line 798": "Não ", "line 799": "Não ", "line 800": "Não ", "line 801": "Não ", "line 802": "Não ", "line 803": "Não ", "line 804": "Não ", "line 805": "Não ", "line 806": "Não ", "line 807": "Não ", "line 808": "Não ", "line 809": "Não ", "line 810": "Não ", "line 811": "Não ", "line 812": "Não ", "line 813": "Não ", "line 814": "Não ", "line 815": "Não ", "line 816": "Não ", "line 817": "Não ", "line 818": "Não ", "line 819": "Não ", "line 820": "Não ", "line 821": "Não ", "line 822": "Não ", "line 823": "Não ", "line 824": "Não ", "line 825": "Não ", "line 826": "Não ", "line 827": "Não ", "line 828": "Não ", "line 829": "Não ", "line 830": "Não ", "line 831": "Não ", "line 832": "Não ", "line 833": "Não ", "line 834": "Não ", "line 835": "Não ", "line 836": "Não ", "line 837": "Não ", "line 838": "Não ", "line 839": "Não ", "line 840": "Não ", "line 841": "Não ", "line 842": "Não ", "line 843": "Não ", "line 844": "Não ", "line 845": "Não ", "line 846": "Não ", "line 847": "Não ", "line 848": "Não ", "line 849": "Não ", "line 850": "Não ", "line 851": "Não ", "line 852": "Não ", "line 853": "Não ", "line 854": "Não ", "line 855": "Não ", "line 856": "Não ", "line 857": "Não ", "line 858": "Não ", "line 859": "Não ", "line 860": "Não ", "line 861": "Não ", "line 862": "Não ", "line 863": "Não ", "line 864": "Não ", "line 865": "Não ", "line 866": "Não ", "line 867": "Não ", "line 868": "Não ", "line 869": "Não ", "line 870": "Não ", "line 871": "Não ", "line 872": "Não ", "line 873": "Não ", "line 874": "Não ", "line 875": "Não ", "line 876": "Não ", "line 877": "Não ", "line 878": "Não ", "line 879": "Não ", "line 880": "Não ", "line 881": "Não ", "line 882": "Não ", "line 883": "Não ", "line 884": "Não ", "line 885": "Não ", "line 886": "Não ", "line 887": "Não ", "line 888": "Não ", "line 889": "Não ", "line 890": "Não ", "line 891": "Não ", "line 892": "Não ", "line 893": "Não ", "line 894": "Não ", "line 895": "Não ", "line 896": "Não ", "line 897": "Não ", "line 898": "Não ", "line 899": "Não ", "line 900": "Não ", "line 901": "Não ", "line 902": "Não ", "line 903": "Não ", "line 904": "Não ", "line 905": "Não ", "line 906": "Não ", "line 907": "Não ", "line 908": "Não ", "line 909": "Não ", "line 910": "Não ", "line 911": "Não ", "line 912": "Não ", "line 913": "Não ", "line 914": "Não ", "line 915": "Não ", "line 916": "Não ", "line 917": "Não ", "line 918": "Não ", "line 919": "Não ", "line 920": "Não ", "line 921": "Não ", "line 922": "Não ", "line 923": "Não ", "line 924": "Não ", "line 925": "Não ", "line 926": "Não ", "line 927": "Não ", "line 928": "Não ", "line 929": "Não ", "line 930": "Não ", "line 931": "Não ", "line 932": "Não ", "line 933": "Não ", "line 934": "Não ", "line 935": "Não ", "line 936": "Não ", "line 937": "Não ", "line 938": "Não ", "line 939": "Não ", "line 940": "Não ", "line 941": "Não ", "line 942": "Não ", "line 943": "Não ", "line 944": "Não ", "line 945": "Não ", "line 946": "Não ", "line 947": "Não ", "line 948": "Não ", "line 949": "Não ", "line 950": "Não ", "line 951": "Não ", "line 952": "Não ", "line 953": "Não ", "line 954": "Não ", "line 955": "Não ", "line 956": "Não ", "line 957": "Não ", "line 958": "Não ", "line 959": "Não ", "line 960": "Não ", "line 961": "Não ", "line 962": "Não ", "line 963": "Não ", "line 964": "Não ", "line 965": "Não ", "line 966": "Não ", "line 967": "Não ", "line 968": "Não ", "line 969": "Não ", "line 970": "Não ", "line 971": "Não ", "line 972": "Não ", "line 973": "Não ", "line 974": "Não ", "line 975": "Não ", "line 976": "Não ", "line 977": "Não ", "line 978": "Não ", "line 979": "Não ", "line 980": "Não ", "line 981": "Não ", "line 982": "Não ", "line 983": "Não ", "line 984": "Não ", "line 985": "Não ", "line 986": "Não ", "line 987": "Não ", "line 988": "Não ", "line 989": "Não ", "line 990": "Não ", "line 991": "Não ", "line 992": "Não ", "line 993": "Não ", "line 994": "Não ", "line 995": "Não ", "line 996": "Não ", "line 997": "Não ", "line 998": "Não ", "line 999": "Não ", "line 1000": "Não ", "line 1001": "Não ", "line 1002": "Não ", "line 1003": "Não ", "line 1004": "Não ", "line 1005": "Não ", "line 1006": "Não ", "line 1007": "Não ", "line 1008": "Não ", "line 1009": "Não ", "line 1010": "Não ", "line 1011": "Não ", "line 1012": "Não ", "line 1013": "Não ", "line 1014": "Não ", "line 1015": "Não ", "line 1016": "Não ", "line 1017": "Não ", "line 1018": "Não ", "line 1019": "Não ", "line 1020": "Não ", "line 1021": "Não ", "line 1022": "Não ", "line 1023": "Não ", "line 1024": "Não ", "line 1025": "Não ", "line 1026": "Não ", "line 1027": "Não ", "line 1028": "Não ", "line 1029": "Não ", "line 1030": "Não ", "line 1031": "Não ", "line 1032": "Não ", "line 1033": "Não ", "line 1034": "Não ", "line 1035": "Não ", "line 1036": "Não ", "line
```

## Preparação para treino

Na língua portuguesa deparamo-nos com vários acentos e outros símbolos como os traços (como exemplo "ensino-te"). Quanto aos acentos, a sua manutenção é imperativa para boa leitura e para remoção de confusões entre palavras (como por exemplo "estás" e "estas"). Por outro lado, os traços foram retirados, já que a flexão do pronome quanto ao número e género é demasiado vasta para garantir um vocabulário conciso por parte do modelo. Outros símbolos de sintaxe e gramaticais foram retirados aquando a preparação para treino.

## 2.3 Modelo

O modelo base foi retirado da plataforma Github, do repositório Seq2Seq-Chatbot [Sojasingarayar, 2020]. Este modelo usa o módulo *tensorflow* para os processos de ML.

No trabalho citado foi também usado um *dataset* de filmes, o 'Cornell Movie Dialog Corpus' [Danescu-Niculescu-Mizil and Lee, 2011]. Este dispunha de 304713 entradas para treino num ficheiro .txt também disponibilizado.

## LSTM

A 'Long Short Term Memory' (LSTM) é uma rede neuronal recorrente capaz de aprender a dependência em problemas de predição. Este é útil em tradução, reconhecimento de fala e, direcionado para este tópico, simulação de conversação humana. Mais sobre este tipo de redes, elas têm um estado interno que consegue representar informação em contexto [Bengio *et al.*, 1994].

## Seq2seq

A metodologia 'sequence-to-sequence' (seq2seq) é a peça-chave do trabalho. A metodologia utiliza modelos de 'Machine Learning' (ML) para obter uma sequência como entrada, num domínio, e convertê-la para uma representação noutra domínio.

O *seq2seq* baseia-se num bloco *encoder* que lê a série de entrada a partir de um vetor de dimensionalidade fixa e um bloco *decoder* que extrai a frase. Ambos estes blocos são células LSTM e são treinados ao mesmo tempo [Sutskever *et al.*, 2014]. Entre eles há um vetor de contexto que encapsula todo o sentido da frase.

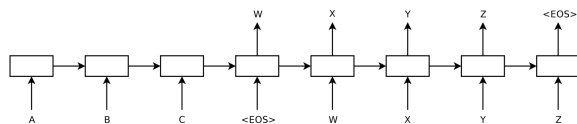


Figure 3: Ilustração da metodologia. os primeiros quatro retângulos referem-se à camada *encoder* enquanto o resto pertence à *decoder* [Sutskever *et al.*, 2014]

No momento inicial, quando é fornecida uma sequência ao modelo, a frase terá de caber no vetor fixo, portanto leva um *padding*. À sequência acrescenta-se a componente "<PAD>". Ao passar pelo vetor, o *decoder* inicia com o comando "<START>", ou "<GO>" como usado neste trabalho, para iniciar a produção da saída. A sequência é completa com a expressão "<EOS>" que simboliza o final do

*output*. Para expressões que o algoritmo não reconhece lança-se o comando "<UNK>".

## 'Attention'

Em ML esta técnica tem o objetivo de reproduzir atenção cognitiva (comportamento humano). o efeito traduz-se em realçar algumas partes da entrada, enquanto diminui outras, mudando o foco da informação dada. Este método é também sensível ao contexto, algo que é tido em conta no treino da máquina.

## 3 Experimentação e Métricas

### 3.1 ROUGE

'Recall-Oriented Understudy for Gisting Evaluation' é um conjunto de métricas especializadas em avaliar *machine translation*. Estas métricas focam-se em quanto as palavras (ou *n-grams* no input se assemelham às previstas pelo modelo.

Neste trabalho foca-se no ROUGE-L que se baseia na subsequência mais longa em comum entre o *output* e a referência. Uma vantagem do uso desta métrica é que não é necessário ter correspondências consecutivas, mas correspondências em sequência [Lin and Och, 2004]. A implementação é feita com recurso ao módulo "rouge" do *python*.

## 4 Resultados

## 5 Problemas e Resoluções

### 5.1 Conjunto de dados

Alguns problemas foram encontrados no uso da informação.

Em primeiro lugar, muitos ficheiros continham referências aos tradutores como parte das legendas. Como estas variavam em formato e local era difícil prever e retirar estas informações com sucesso.

Algumas traduções eram também desvirtuadas do português Portugal, tais como o uso de expressões como "em a", em vez de "na" em vários locais. A fonte oferecia também um pacote de português Brasil, portanto estas nuances não se esperariam no conjunto usado. Pequenas alterações como esta seriam de difícil execução sem uma análise extensiva dos dados.

### Resolução

Apesar destes problemas, usou-se este 'dataset' pela sua vastidão e facilidade de acesso. Acrescendo a dificuldade de materiais para a língua desejada, esta foi a melhor opção encontrada.

### 5.2 Modelo

A escolha do código foi influenciada na metodologia seq2seq (aconselhada pelo professor). A pesquisa foi então restringida para tal.

O uso desta versão 1.14 do *tensorflow* é desatualizada, pois o pacote está já na versão 2 a partir de 2019. Esta veio com uma mudança na organização e, portanto, na sintaxe a usar para construir o modelo. Obriga, também, ao uso de uma versão anterior do *python* (3.6 ou menor).

Este modelo traz então algumas agravantes, em especial o tempo de compilação. Para 500 filmes com 100 *epochs*, o

tempo estimado era de cerca de 13 horas, todo o processo de treino. Com estes números o modelo leva a um esforço computacional aquém do esperado, tornando-se até inexecutável em termos de desenvolvimento do trabalho.

Guardar os pesos ocupa muito espaço!!!!?

## 6 Conclusions

### References

- [Bengio *et al.*, 1994] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.
- [Danescu-Niculescu-Mizil and Lee, 2011] Cristian Danescu-Niculescu-Mizil and Lillian Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*, 2011.
- [Lin and Och, 2004] Chin-Yew Lin and FJ Och. Looking for a few good metrics: Rouge and its evaluation. In *Ntcir workshop*, 2004.
- [opensubtitles, 2022] opensubtitles. opensubtitles.org, 2022.
- [OPUS, 2022] OPUS. OPUS the open parallel corpus, 2022.
- [Sojasingarayar, 2020] Abonia Sojasingarayar. Seq2seq-chatbot. <https://github.com/Abonia1/Seq2Seq-Chatbot>, 2020.
- [Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.