Neural Machine Translation using LSTM with MHA and food-related ontologies

Liviu-Ștefan Neacșu-Miclea – ICA 256/2

The Ontologies

ΑΜΑΛΘΕΙΑ (Amaltheia)

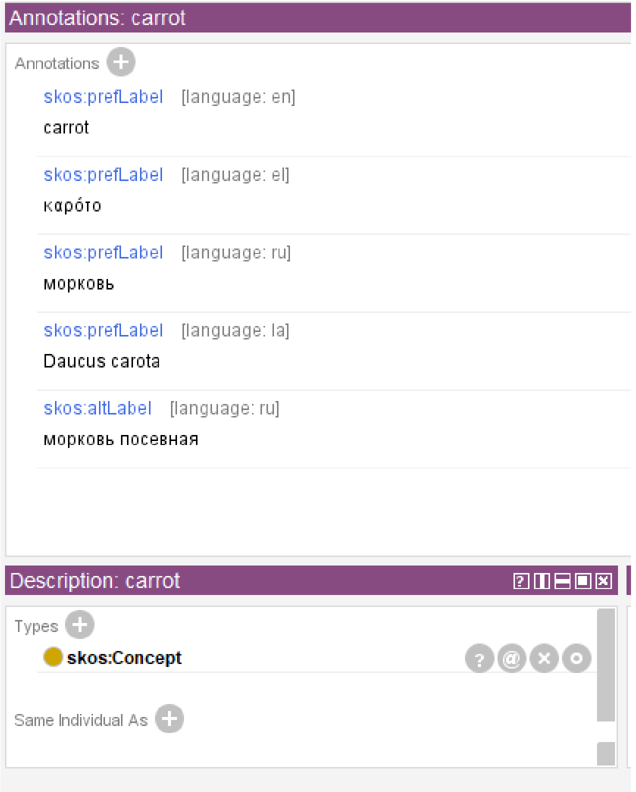
- Has a single Concept class (no hierarchy)
- Has annotated multilingual labels for each individual

FoodOn

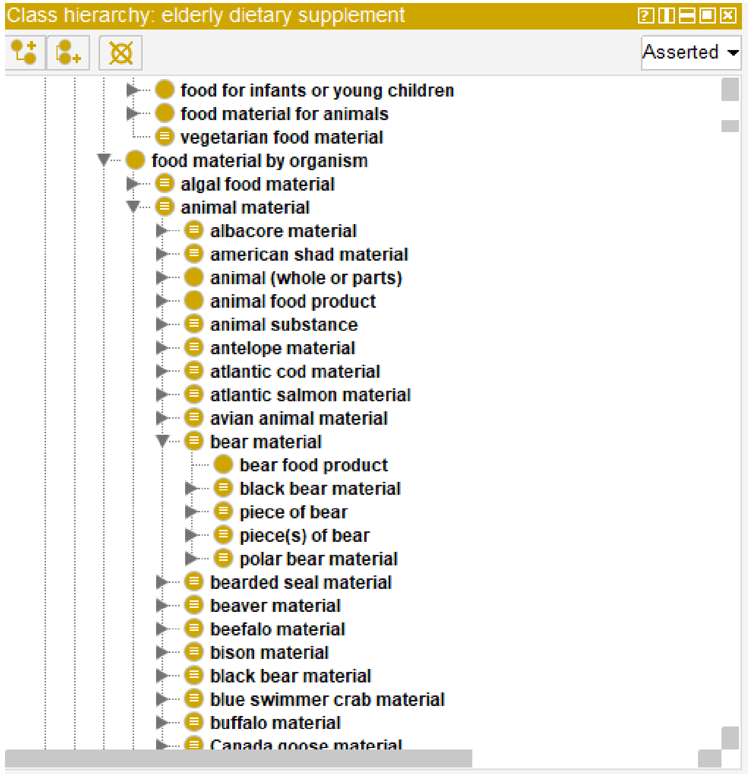
- Has very detailed gastronomical hierarchy
- No multilingual labels
- No individuals except for country names and units of measure

The Ontologies

ΑΜΑΛΘΕΙΑ (Amaltheia)



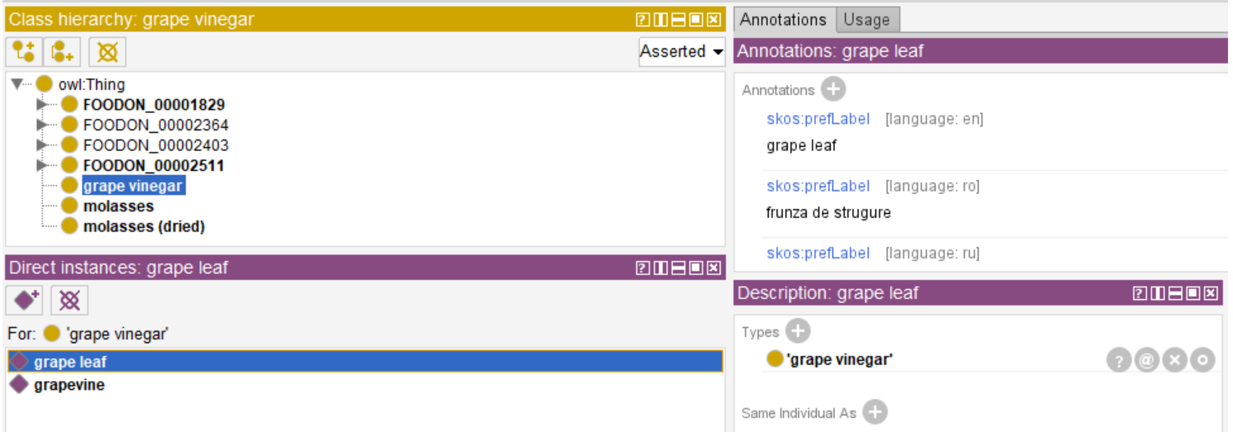
FoodOn



The Ontologies

ΑΜΑΛΘΕΙΑ (Amaltheia) + FoodOn = mini_food_ontology

- I combined (intersected) the two ontologies
- Keep FoodOn's **hierarchy**
- Keep Amaltheia's **multilingual labels**
- Use google translate API to add Romanian translation labels automatically for individual concepts



The Corpus

- OPUS: WikiMatrix.en-ro parallel corpus (opus.nlpl.eu)

1	The first was: "Are there many Jews who think like you?"	Primul a fost: „Sunt mulți evrei care gândesc ca dumneavoastră?”
2	(And I recite to you Allah's statement:) “O People of the Scriptures!	O, oameni ai Cărții!
3	His are the heavens and the earth.	Iată istoria cerurilor și a pământului, când au fost făcute.
4	16 And many of the children of Israel shall he turn to the Lord their God.	El va întoarce pe mulți din fiii lui Israel la Domnul, Dumnezeul lor.
5	How shall I have a son when no man hath touched me?”	Cum să am un copil, când nici un bărbat nu m-a atins?”
6	Anyone who tells you differently is a liar."	Oricine vă spune altfel e un mincinos".
7	In return he would forget Orton ever existed.	În schimb ar uita că Orton a existat vreodată.
8	Q: How can you stop an Albanian tank?	Î: Cum poți opri un tanc albanez?
9	Jack London is not mentioned.	Jack London nu este menționat.
10	Why is there no fish on the market?	De ce nu există pește pe piață?
11	K. visits the lawyer several times.	K. vizitează avocatul de mai multe ori.
12	But Francis and I worked together ...	Însă Francis și cu mine am lucrat împreună ...

Text Preprocessing

Mask foreign characters
Keep latin, Romanian diacritics,
Some common symbols
Mask everything with #

Fit a 1024 wordpieces model
common for both languages

Filter → Cleanup → Annotate → Wordpieces

Remove sentences
Unrelated to ontology
(fuzzy matching)

Fuzzy match ontology
Concepts again and
mark them in-text

Totalling 64084 English-Romanian paired samples

Text Preprocessing

Filtered

Eddy eventually succumbed to his hunger and ate human flesh, but that was soon gone.
Eddy până la urmă a cedat și a mâncat și el carne umană, care curând s-a terminat.

Cleaned

eddy eventually succumbed to his hunger and ate human flesh, but that was soon gone.
eddy până la urmă a cedat și a mâncat și el carne umană, care curând s-a terminat.

Annotated

eddy eventually succumbed to his hunger and ate human [FOODON_33]flesh, but that was soon gone.
eddy până la urmă a cedat și a mâncat și el [FOODON_33]carne umană, care curând s-a terminat.

But...

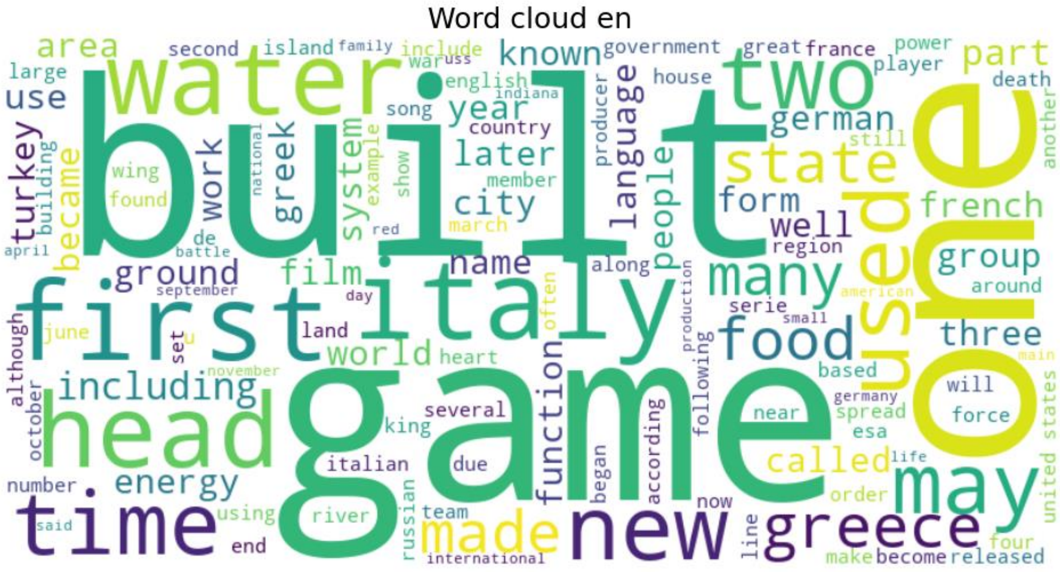
Mismatched translations

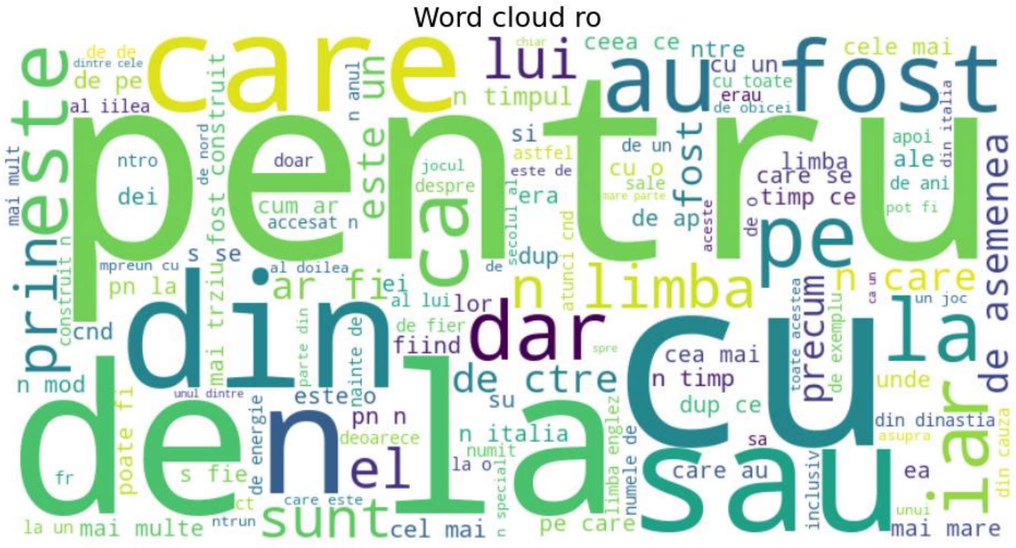
The Italian word frittata derives from friggere and roughly means fried.
cuvântul [FOODON_394]musaca vine din arabă și înseamnă servit rece.

Wrong annotations (Omonyms)

după ce s-a lăsat cortina de [FOODON_1300]fier, ...

Text Preprocessing





Text Preprocessing

0 \n	103 ia	206 rea	309 fer	412 ase	515 war	618 cun	721 used	824 ține	927 locu
1 \s	104 pe	207 ag	310 da	413 cum	516 public	619 int	722 way	825 indu	928 work
2 !	105 cu	208 io	311 cor	414 led	517 dou	620 fel	723 vă	826 ence	929 well
3 "	106 ru	209 ame	312 nor	415 cel	518 has	621 spre	724 ited	827 comple	930 har
4 #	107 ol	210 bu	313 ated	416 ară	519 hi	622 some	725 count	828 him	931 icul
5 %	108 ma	211 ital	314 wn	417 ear	520 ations	623 vis	726 brit	829 vit	932 dus
6 &	109 ra	212 pă	315 wo	418 low	521 after	624 dre	727 alb	830 ely	933 prima
7 '	110 el	213 ile	316 sau	419 et	522 tul	625 rou	728 france	831 tate	934 thern
8 (	111 con	214 ele	317 bl	420 med	523 tele	626 dintre	729 fiind	832 ps	935 ise
9)	112 ct	215 pri	318 cer	421 după	524 fă	627 trans	730 dy	833 atul	936 bon
10 ,	113 au	216 sa	319 pol	422 eng	525 mis	628 ica	731 food	834 abo	937 gi
11 -	114 ui	217 that	320 ard	423 ding	526 lea	629 mer	732 kno	835 van	938 ade
12 .	115 ate	218 ber	321 pul	424 ște	527 gree	630 world	733 func	836 mal	939 ange
13 /	116 ii	219 sc	322 inter	425 ry	528 țin	631 mbrie	734 aceast	837 sin	940 qui
14 0	117 din	220 ite	323 wor	426 cat	529 roman	632 sco	735 viz	838 arch	941 known
15 1	118 us	221 jo	324 cent	427 tin	530 ong	633 ory	736 char	839 timpul	942 util
16 2	119 me	222 ră	325 can	428 water	531 vie	634 bal	737 able	840 contro	943 inst
17 3	120 ad	223 area	326 cal	429 les	532 sing	635 bra	738 duce	841 came	944 pun
18 4	121 ve	224 ke	327 dis	430 fil	533 când	636 către	739 fam	842 any	945 star
19 5	122 ap	225 av	328 mi	431 put	534 joc	637 bul	740 cle	843 while	946 ație
20 6	123 ut	226 nu	329 port	432 iar	535 elor	638 ha	741 aliz	844 scri	947 maced
21 7	124 lo	227 ment	330 which	433 ical	536 sup	639 cond	742 ament	845 greece	948 toare
22 8	125 for	228 gre	331 zi	434 est	537 tem	640 rest	743 ura	846 ses	949 ough
23 9	126 po	229 pl	332 ces	435 had	538 tal	641 ka	744 lit	847 famil	950 iile
24 :	127 fo	230 min	333 son	436 ass	539 sho	642 bet	745 ape	848 asemenea	951 ției
25 ;	128 pro	231 ții	334 ten	437 first	540 stat	643 rile	746 mas	849 războ	952 ide
26 ?	129 ste	232 tă	335 game	438 euro	541 elect	644 ior	747 main	850 rez	953 then
27 `	130 be	233 tim	336 up	439 built	542 ani	645 ions	748 tit	851 tic	954 arte
28 a	131 di	234 ând	337 ave	440 gan	543 var	646 organ	749 my	852 city	955 ctions
29 b	132 du	235 por	338 rit	441 oc	544 out	647 europe	750 asem	853 dio	956 dan
30 c	133 că	236 ther	339 ure	442 new	545 ata	648 over	751 ună	854 cunos	957 dat

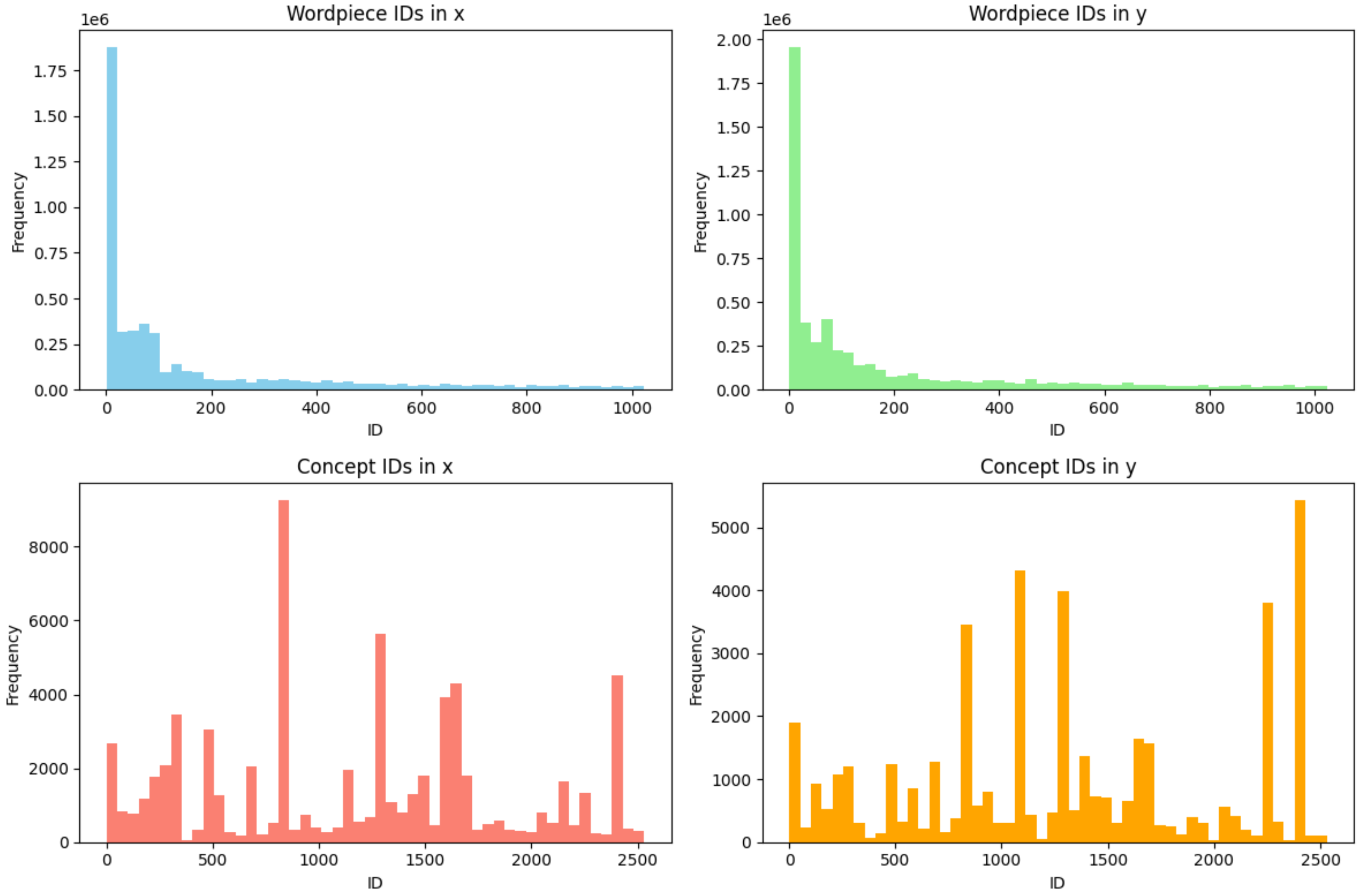
Example wordpieces

Text Preprocessing

- Finally, unique ids were assigned to each individual wordpiece (‘w:<id>’) or concept (‘c:<id>’)

- There were too many concepts in mini food ontology => only the top 128 most frequent were chosen for train

- Max sentence length in ids: 256 (accounts for >95% of all corpus, because of long tail distribution of sentence length)
- Smaller lengths => pad till 256 with w:0 (aka \n)

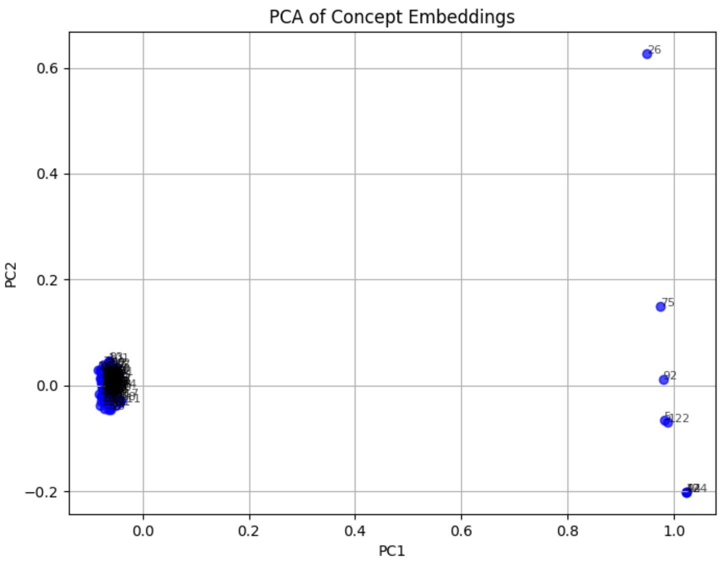


Embeddings

Vectorized tokens, length 127 + 1 dimension for type (-1 concept, 1 wordpiece)

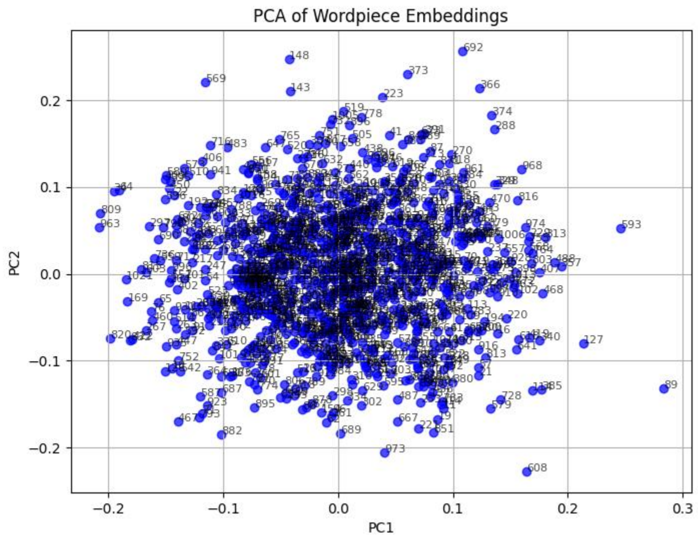
Concept Embeddings

- Poincaré projection computed on hierarchy concept graph
- Hyperbolic to Euclidean conversion



Wordpiece Embeddings

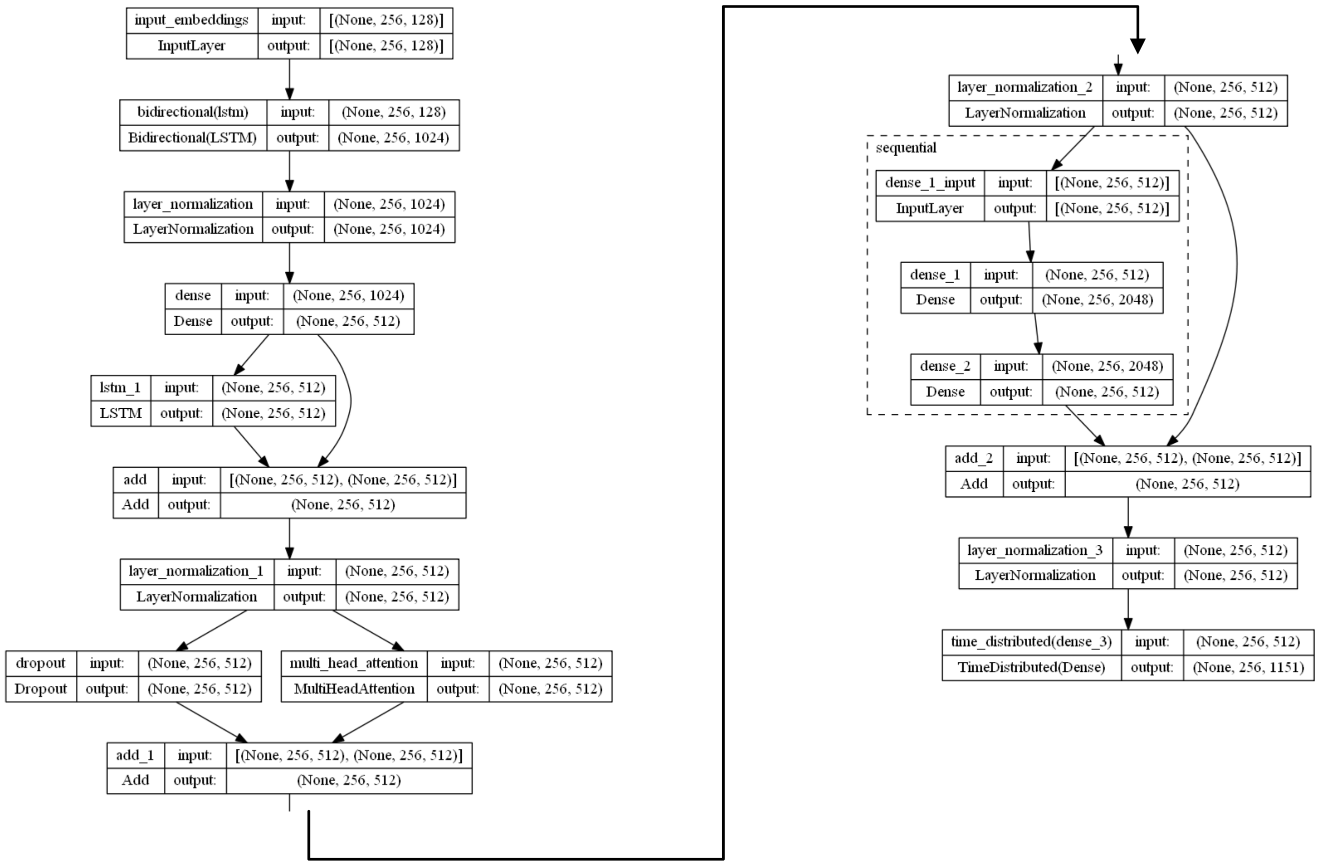
- Randomly chosen from the vector space



NMT Architecture

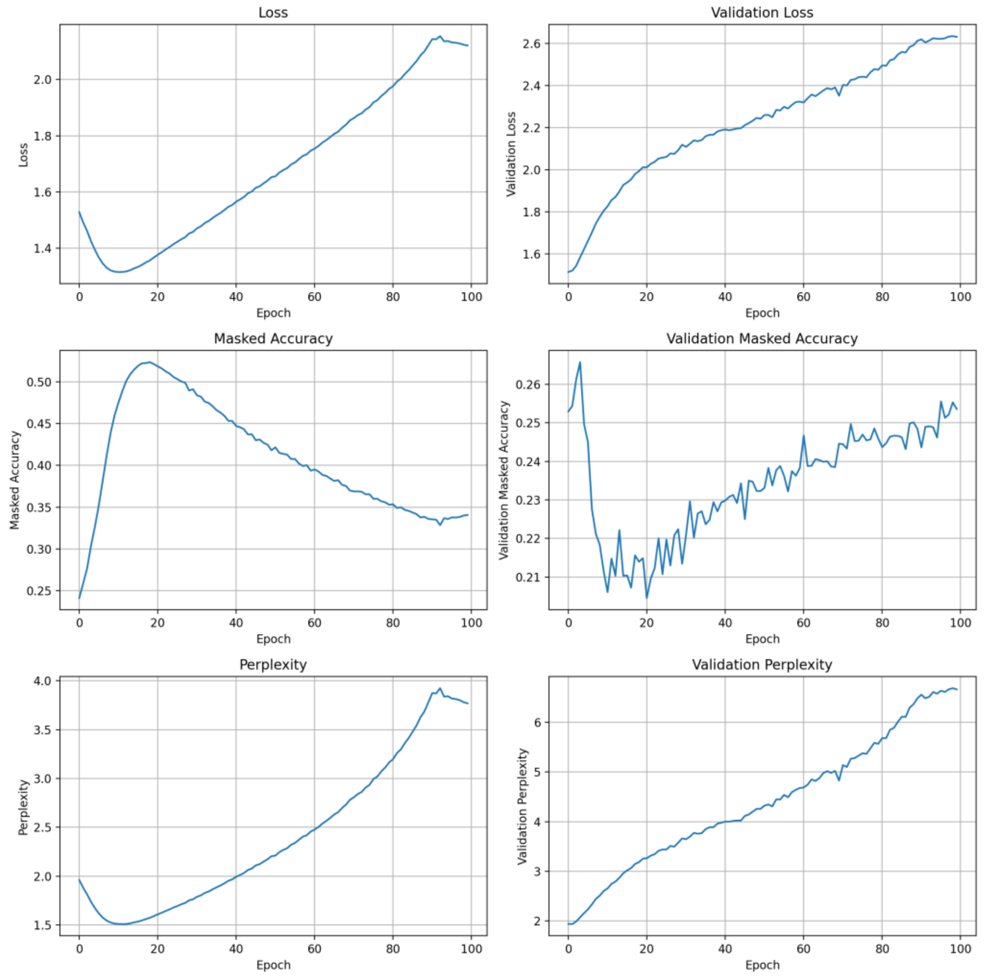
- Bi-LSTM
- LN+Dense
- LSTM+Skip connection
- LN
- MHA
- Fully-Connected
- LN
- Softmax (concepts+wordpieces concatenated categoricals)

- **Curriculum learning**
  - Promote shorter sentences in the early epochs
- Loss = weighted CCE & Focal loss



Experiments

Metric	Value (Val data)
BLEU-4	0.0124
ROUGE-2 ROUGE-L Accuracy	0.0487 0.3298 0.256



Experiments

```
en =  i am eating a delicious [FOODON_939]ice cream.
ro =  ti  tre   o       poant.

en =  the cake contains strawberries.
ro =  aregenareare areun ere   e ați..

en =  it is distinct from the [FOODON_74]mustard plants which belong to the genus brassica.
ro =  se dist men ș plande de șștar carecareapar eeului brassica

en =  saturn is the sixth planet from the sun and the second-largest in the solar system, after jupiter.
ro =  saturn este dea  dinta en   esuleses     li do  mare ile ile ile  l gagaistem

en =  it is used in its [FOODON_872]dried form for japanese soups, tempura, and material for manufacturing [FOODON_872]d
ried nori and tsukudani and [FOODON_1627]rice.
ro =  este folosfolos               ja ja,,        ,                    uu
```

```
[EN] >> I am eating bread.
en =  i am eating [FOODON_93]bread.
ro =  tou suntpapachemâncatineă.
```

Discussion

- Poor performance because
  - The complexity of the task in general

  - Potential noise in data (invalid translations, bad annotations)

  - Unmeaningful embeddings
    - Poincare concept embeddings tend to cluster near a point due to projection of hyperbolic onto Euclidean
    - Wordiece embedding were just randomly chosen

  - Weak architecture (compared to GNMT for example)

Future work

- Carefully clean the train data
- Improve embeddings
  - Try more sparse concept emeddings
  - Word2vec-like on wordpieces
- Improve training pipeline (e.g. losses with length masks & weights)
- Try better architectures

Thank you!