☆HW 2

- Word Sense Disambiguation: computational identification of the meaning of the words in context.

OSS: can be seen as a multi-class classification problem where the classes belong to a specific Sense inventory

- Task challenges

1) Sense distribution follow the Zipf distribution => class highly unbalanced
2) Number of possible senses in the hundred of thousands
3) Competence necessaria per annotare correttamente i significati => difficile scalare su diversi domini e languages
4) Polysemous words have senses with similar meanings => the model struggle to classify them correctly

Def:
- Polysemous: polysemous words are those that, based on context, can assume different senses
- Homonyms: two words written in the

some way, but they're different with different meanings

OSS's

La differenza sostanziale tra omonimia e polisemia consiste nel fatto che **le parole omonime sono parole diverse**, con radici completamente differenti e differenti significati. Le **parole polisemiche** sono invece **parole uniche che nell'uso comune assumono molteplici significati**.

- WSD with coarse grained senses

We can cluster the meanings in groups where, for each group, we keep the most similar meanings

Polysemy vs Homonymy

Idea: cluster the similar senses of a polysemous word to obtain a list of **highly distinguishable, coarse grained candidates.**

Two or more words are **homonyms** when they have the same lexical form but different, **unrelated** meanings.

- **Italic.n.01:** a style of handwriting with the letters slanting to the right.
- **Italic.n.03:** a typeface with letters slanting upward to the right.

Homonymy

- **Italic.n.02:** a branch of the Indo-European languages of which Latin is the chief representative.

SAPIENZA NLP

this simplify the Sch because now we have a clear distinction between the classes

Coarse-Grained WSD

WSD with coarse-grained senses

Instead of using all the candidates, now we have classes that are highly distinguishable: we have clustered **candidates with similar meaning** in ***homonymy clusters***.

WordNet (Miller, 1995)

- **Italic.n.h.01:** {a style of handwriting with the letters slanting to the right; a typeface with letters slanting upward to the right }.
- **Italic.n.h.02:** {a branch of the Indo-European languages of which Latin is the chief representative.}

In this way we have to disambiguate only between homonymy clusters, which is easier due to their **distant meanings**.

Def: • Wordnet: a lexical-semantic database containing structured knowledge for the english language.

Coarse-Grained WSD

WSD with coarse-grained senses

WORD WORD SENSE

WordNet (Miller, 1995)

- **race.n.01**: any competition.
- **race.n.02**: a contest of speed.
- **race.n.03**: people who are believed to belong to the same genetic stock.
- **subspecies.n.01**: (biology) a taxonomic group that is a division of a species.
- **slipstream.n.01**: the flow of air that is driven backwards by an aircraft propeller.
- **raceway.n.01**: a canal for a current of water.

↓

fine-to-coarse mapping to **Homonymy Clusters**

↓

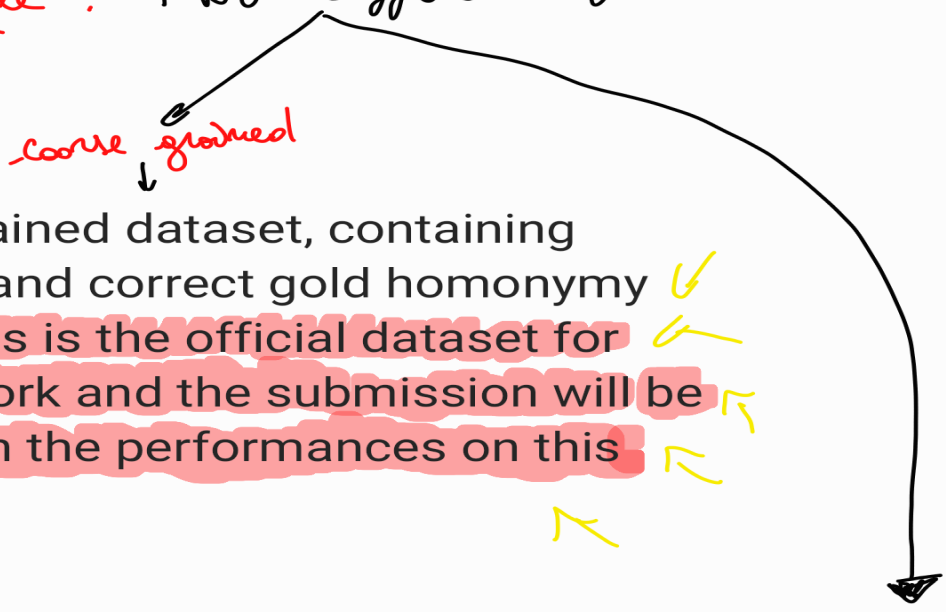
race.n.h.01
- **race.n.01**: any competition.
- **race.n.02**: a contest of speed.
- **slipstream.n.01**: the flow of air that is driven backwards by an aircraft propeller.
- **raceway.n.01**: a canal for a current of water.

race.n.h.02
- **race.n.03**: people who are believed to belong to the same genetic stock.
- **subspecies.n.01**: (biology) a taxonomic group that is a division of a species.

- Dataset → Wordnet

• Samples: Each sample is a tokenized sentence with the information about the instances to disambiguate (es. POS tag)

• Dataset file: Two different files



A coarse-grained dataset, containing candidates and correct gold homonymy clusters. This is the official dataset for this homework and the submission will be evaluated on the performances on this data.

A fine-grained dataset, containing candidates and correct gold WordNet senses. This is a key resource for you to obtain bonus points by doing novel comparative

PS: There's a 3-rd file:

You will receive an additional file, "coarse_to_fine.json" containing a mapping between each coarse grained candidate and its fine-grained sub-senses along with their definitions.

• Focus on dataset structure + coarse-to-fine

1) The Dataset

Coarse-grained dataset (mandatory usage)

- Each sample of the coarse-grained dataset is a sentence with annotations about words and their senses:
  - **idx**: document id and sentence id
  - **instance_ids**: mapping between token based offsets of each instance and its id
  - **words**: list of tokenized words
  - **lemmas**: list of tokenized and lemmatized words
  - **pos_tags**: list of part of speech tags
  - **senses**: mapping between token based offsets and gold homonymy clusters
  - **candidates**: list of possible homonymy clusters for each instance

The Dataset

Data Format: coarse-grained dataset

```
{
    "d000.s002": {
        "instance_ids":{
            "1": "d000.s002.t000"
            "5": "d000.s002.t001"
        }
        "lemmas": ["the", "race", "will", "take", "place", "today"],
        "words": ["The", "races", "will", "take", "place", "today"]
        "pos_tags": ["DT", "NOUN", "VB", "VB", "NOUN", "ADP"]
        "senses": {
            "1": "race.n.h.01"
            "5": "today.r.h.01"
        }
        "candidates":[
            "1": ["race.n.h.01", "race.n.h.02"]
            "5": ["today.r.h.01"]
        ]
    },
```

DOCUMENT ID · SENTENCE ID · OFFSET DELLA PAROLA NELLA FRASE · ID · 1 · 2 · homonimy cluster label · all senses of 1

Quindi se ho capito bene **senses** è la golden label, candidates sono tutti i possibili significati delle parole da disambiguare => in evaluation è corretto se senses = candidates

2)

The Dataset

Fine-grained dataset (recommended for bonus points)

- Each sample of the fine-grained dataset is a sentence with annotations about words and their senses:
  - **idx**: document id and sentence id
  - **instance_ids**: mapping between token based offsets of each instance and its id
  - **words**: list of tokenized words
  - **lemmas**: list of tokenized and lemmatized words
  - **pos_tags**: list of part of speech tags
  - **senses**: mapping between token based offsets and gold WordNet synsets
  - **candidates**: list of possible WordNet synsets for each instance

The Dataset

Data Format:fine-grained dataset

```
{
    "d000.s002": {
        "instance_ids":{
            "1": "d000.s002.t000"
            "5": "d000.s002.t001"
        }
        "lemmas": ["the", "race", "will", "take", "place", "today"
        "words": ["The", "races", "will", "take", "place", "today"
        "pos_tags": ["DT", "NOUN", "VB", "VB", "NOUN", "ADP"]
        "senses": {
            "1": "race.n.02"
            "5": "today.r.01"
        }
        "candidates":[
            "1": ["race.n.01", "race.n.02", "race.n.03",
"subspecies.n.01", "slipstream.n.01", "raceway.n.01"]
            "5": ["today.r.01", "today.r.02"]
        ]
    },
```

3)

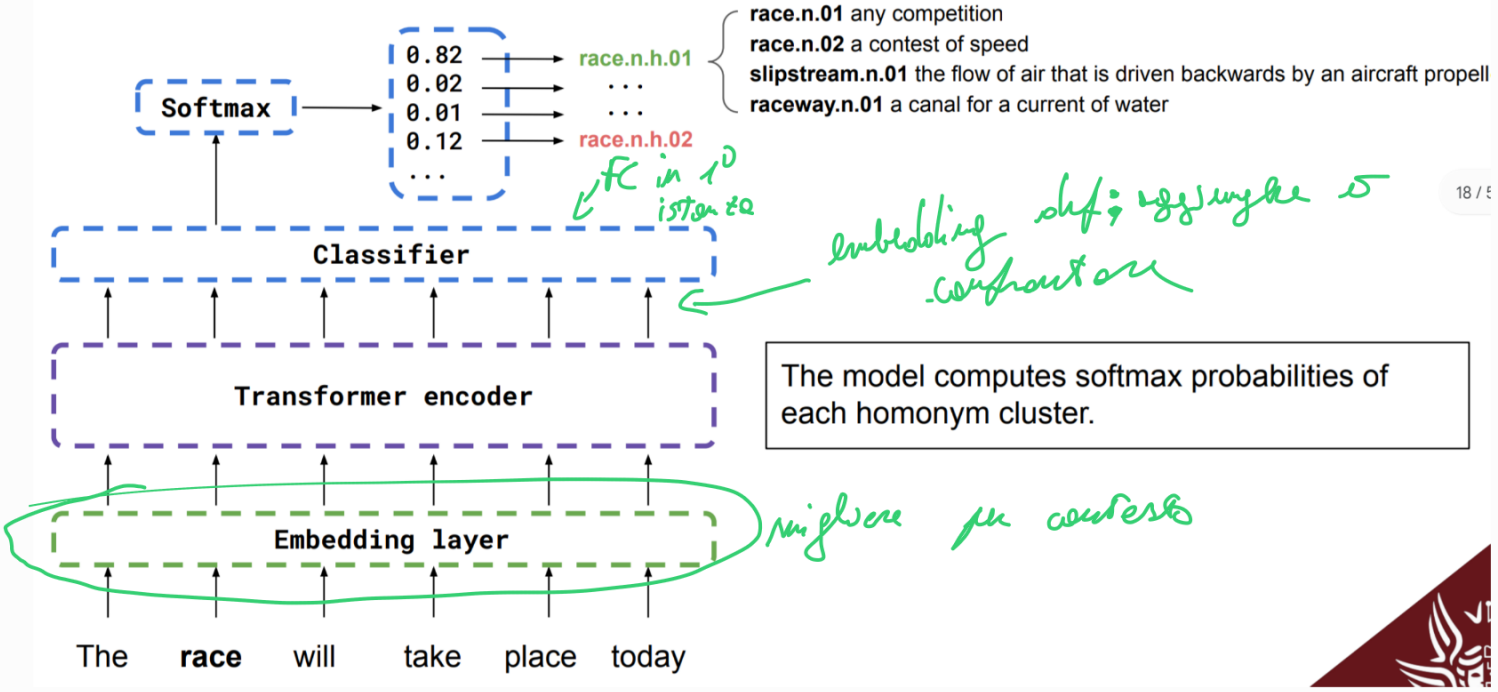
Additional data

Coarse-to-fine mapping

```
{"race.n.h.01": [
      "race.n.01" : "any competition",
      "race.n.02" : "a contest of speed",
      "slipstream.n.01" : "the flow of air that is driven backwards by an aircraft
propeller",
      "raceway.n.01" : "a canal for a current of water"
],
"race.n.h.02": [
      "race.n.03" : "people who are believed to belong to the same genetic stock",
      "subspecies.n.01" : "(biology) a taxonomic group that is a division of a
species"
],
"today.r.h.01": [
      "today.r.01" : "in these times"
      ...
],
}
```

- Possible approach

Coarse-Grained WSD as Multiclass Token Classification



- Extras:

Increase the complexity of your model!

- Take inspiration from recent papers:
  - **GlossBERT**: BERT for word sense disambiguation with gloss knowledge (ACL 2019)
  - **EWISER**, Breaking Through the 80% Glass Ceiling: Raising the State of the [illegible] in Word Sense Disambiguation by Incorporating Knowledge Graph Information(ACL 2020)
  - **BEM**, Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders (ACL 2020)
  - **ESC**: Redesigning WSD with Extractive Sense Comprehension(NAACL 2021)
  - **ConSeC**: Word Sense Disambiguation as Continuous Sense Comprehension (EMNLP2021)

Fine- vs. coarse-grained WSD

- Using the **fine-grained** version of the dataset you can train a standard WSD model:
  - **Compare the results** of your architectures on the two tasks.
  - **Apply the coarse-to-fine mapping to the output** of the fine-grained wsd model to obtain coarse-grained disambiguations.
    - Is it better than the model trained on coarse-grained data? (If motivated on your report, you can submit this model!)
  - **Use both systems**: you can use the coarse grained system to filter out senses for the fine-grained wsd.
  - Use **latent homonymy cluster embedding** to add useful information.
  - WSD coarse-grained by training a fine grained system that is rewarded positively for every synset in the correct homonym
  - Train a multiclass multilabel classifier (multilabel for each sense in a given homonymy cluster)
  - **Analyze qualitatively your results.**

Other extras

- Use sense definitions:
  - Find a way to employ senses definitions in your pipeline (it will improve your results!)
- Find new homonyms: during test?
  - Find ways to detect new homonyms and validate their contribution
- Test on Multilingual Dataset:
  - Building on a multilingual homonym detector, you can train a Multilingual Model on coarse grained data, and test if the model is able to generalize well in other languages.
  - Possible Multilingual resources:
    - XL-WSD: An Extra-Large and Cross-Lingual Evaluation Framework for Word Sense Disambiguation.
- Extend coarse granularity to Entities: ?
  - Coarse Grained named entity recognition
  - Possible start of a thesis project with SapienzaNLP!

- To use:

1) Pytorch lightning
2) Allen Nlp (?)
3) Torchtext
4) NLTK

} frameworks

5) Word 2 Vec, GLOVE (Word embedding)
6) Contextualized word embeddings (Elmo)
7) Transformer-based models (Bert, BART, ROBERTa)
   ↳ suggested

---

1) Vedere come apprendere al meglio il contesto della frase (probabilmente transformers)

2) Utilizzare le definizioni dei significati come "contesto" per la scelta del significato corretto.