Assignment 2 - Semantic Role Labelling

1. Introduction

- Please read the document very carefully. If you have questions ask in the telegram group of the course (link can be found in Canvas).

- Use this template to complete your assignment and upload it to Canvas: https://colab.research.google.com/drive/1ttPT6X4K0ovgbzmNjlcEiprkj1LaBuF2#scrollTo=8a46ab45-d215-41af-b910-63ff4a215a07

- Submit your results to Codalab (both Practice and Test tracks): https://codalab.lisn.upsaclay.fr/competitions/531
    - Important: indicate the team name “Intro to NLP 2023”.

2. Task Description

2.1 Introduction

Our task is motivated by human need to compare various objects: different models of mobile phones, cars, programming languages, countries, etc. This information need has been addressed in NLP research, but there is currently much room for the improvement of existing comparative question answering systems.

For example, the CAM (Comparative Argument Mining) system[1] is given a pair of objects to compare and retrieves arguments in favour each of them. It also extracts **predicates** (comparative characteristics of the objects, e.g. *easier*, *better*, *faster*, etc.) and **aspects** (features along which the objects are compared, e.g. *speed*, *screen*, *performance*, etc.) from the arguments (comparative sentences).
Aspects and predicates are extracted using hand-written templates which have low recall (fail to extract entities which do not conform to the templates) and occasionally extracts incorrect entities.

[1] http://ltdemos.informatik.uni-hamburg.de/cam/

We would like to improve its performance by training a model which extracts objects, aspects, and predicates from a sentence. This model should be trained on sentences where words or phrases are labelled with entities.

Examples of sentences

Postgres is easier to install and maintain than Oracle.

[Postgres OBJECT] is [easier PREDICATE] to [install ASPECT] and [maintain ASPECT] than [Oracle OBJECT].

Instances can be multiword:

Advil works better for body aches and pains than Motrin.

[Advil OBJECT] works [better PREDICATE] for [body aches ASPECT] and [pains ASPECT] than [Motrin OBJECT].

Data format

The provided data files are in CoNLL format. Each line contains one word and its label, separated by a tab ("Word<TAB>label"), the end of the sentence is marked with an empty line. The labels are in BIO format, where each of the entity labels ("Object", "Aspect", "Predicate") is prepended with a prefix "B-" or "I-", indicating the **b**eginning of an entity (the first word of an entity) and the **i**nside of an entity (the second and all subsequent words). Words which are not a part of an entity are labelled with "O":

advil B-Object

works O

better B-Predicate

for O

body B-Aspect

aches I-Aspect

and O

pains B-Aspect

than O

motrin B-Object

. O

2.2 Task formulation

The data consists of comparative sentences (i.e., sentences which contain comparison of two or more objects). The data contains three types of entities:

- **Object** - objects which are being compared
- **Aspect** - features along which the objects are compared
- **Predicate** - words or phrases which implement the comparison (usually comparative adjectives or adverbs)

The dataset uses BIO labelling scheme:

- The first word of an entity is labelled with "B-<entity-type>" (*beginning* of an entity)
- The second and further words of an entity are labelled with "I-<entity-type>" (*inside* of an entity)
- Words which are not a part of an entity are labelled with "O" (*out* of entity)

Therefore, our dataset uses the following labels:

- O
- B-Object
- I-Object
- B-Aspect
- I-Aspect
- B-Predicate
- I-Predicate

Your task is to assign one of such labels to each of the words in the test set.

2.1.3 Evaluation metrics

The result will be evaluated with F_1-score:

$$2 \cdot \frac{precision \cdot recall}{precision + recall} = 2 \cdot \frac{tp}{tp + 0.5(fp + fn)}$$

We will consider scores for all individual classes except O. For multi-word entities we use the "relaxed" metric: if the borders of a predicted entity match those of the reference entity, we add 1 to *tp* quantity (number of *true positive* examples). If there is only a partial match, we add a number between 0 and 1 computed as the intersection length divided by the full entity length.

2.1.4 Method

Your task is to train a sequence labelling model on a provided labelled dataset. You can use neural architectures (LSTM, Transformer) as well as traditional ML algorithms for sequence labelling (CRF, HMM). We encourage you to experiment with different types

of embeddings (e.g. context-free GloVe and fastText or context-informed ELMo and BERT).

In the context of this assignment, you will solve a sequence labelling task on the dataset of comparative sentences provided by the course team. You will need to train a model and submit your solution to the CodaLab competition.

Related to this seminar are materials of the lecture on sequence tagging. You can try adopt solution from the seminar to solve the suggested task and evaluate it. Modification of the solution may be done though:

- Testing more recent pre-trained models, larger base models.
- Adding more features.
- Using multi-task learning i.e., pre-training sequence taggers on related sequence tagging datasets to boost performance on the target task, e.g. on NER or SRL. You can find the related datasets in the open sources and use them. It's desirable to perform some kind of ablation study to check if the additional data harm or help performance.