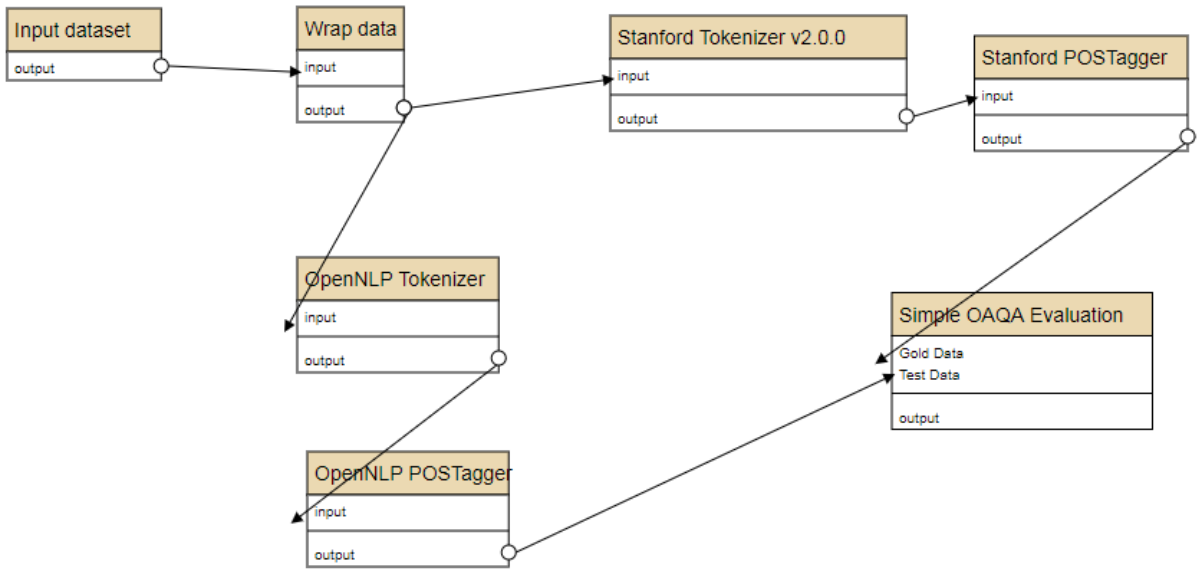
Behaviour of POS Tagging in LAPPS - Results Report

The goal is to investigate the difference between different existing POS Tagger tools in LAPPS of **Stanford POS Tagger** and **OpenNLP POS Tagger**. This assessment is performed on the **Final Diagnosis** section of pathology reports.

The two workflows are run in parallel for each POS Tagger tool. Each workflow uses a different tokenizer which is likely to contribute to differences in results for POS Tagging. The Stanford POS Tagger used the corresponding Stanford Tokenizer, while the OpenNLP POS Tagger used the results of the OpenNLP Tokenizer.

The workflow in LAPPS for such a POS comparison is shown below.



The workflows were applied 20 different pathology reports and then a report presents the difference in statistics both overall and for each single file.

The results show the most common tokens that are tagged differently. The following table shows the token and POS tags from the two analyses, as well as the number of times such differences occur in all 20 pathology reports. For each entry of the table below, the POS tag which we judge to be correct is highlighted in red.

Some qualitative analysis has been done on the differences table, which gives some insights into the characteristics and drawbacks of the two tokenization and POS tagging system. The most important findings are presented in the following points:

- The most common difference occurs around the section title “Final Diagnosis”. The differences are caused by the presence of several newline symbols “\n” around the title. The Stanford POS Tagger splits the string “\n\nFinal” into a tokens based on the “\”, and therefore “\”, “n”, and “nFinal” are tokenized as separate tokens. However, OpenNLP POS Tagger splits the tokens by spaces, and therefore the whole string “\n\n\nFinal” is tokenized into a single token.
- Another critical issue is that usually the word immediately after the section title “Final Diagnosis” could not be tokenized correctly since there are newline symbols between them. For example, “Diagnosis\n\nColonic” is either tokenized into a single token or “nColonic” as a token. We treat this as an important point since in some cases the word right after the section title has some direct indication about the cancer case.
- There are in total 13 instances where the word “Final” is tagged as JJ(adjective) in Stanford POS Tagger and NN(Noun) in OpenNLP POS Tagger. In 11 cases of these differences, the Stanford POS Tagger is correct so the word should be tagged as JJ.
- There are 8 instances where the word is tagged as NN in Stanford POS Tagger and NNP in OpenNLP POS Tagger. Stanford POS Tagger is correct in 1 case and OpenNLP POS Tagger is correct in 7 cases.

From these differences we have identified improvements to the tagging and tokenization systems to solve these issues. Our implementation for improvements include:

- Assessing the severity of different tag confusions before choosing the POS Tagger system to use. For example, we may treat confusion between JJ and NN more severely than NN and NNP. Therefore, in that case, Stanford POS Tagger should be used since it is more often correct.
- Adding a post-processing stage after tokenization to deal with the newline symbol. Whenever we find the symbol “\n” in the original text and it is followed by a word starting with capital letter, we tag this symbol as a newline symbol first and split the tokens around it before going into the POS tagging system. This same strategy needs to be applied to other ESC characters.

Table 1: Summary of POS Tag differences for 20 files

Stanford POS Tagger		OpenNLP POS Tagger		Number of times
POS	Text	POS	Text	
NN	\	JJ	\n\n\nFinal	22
NN	n			
CD	\			
NN	n			
CD	\			
JJ	nFinal			
NN	Diagnosis	NNP	Diagnosis\n\nColonic	3
NN	\			
NN	n			
VBG	\			
JJ	nColonic			
NN	Diagnosis	NNP	Diagnosis\n\nFragments	1

NN	\			
NN	n			
VBG	\			
JJ	nFragments			
JJ	intramucosal	NN	intramucosal	4
VBN	differentiated	JJ	differentiated	6
JJ	villotubar	NN	villotubar	1
NN	Diagnosis	NNP	Diagnosis\n\nPoorly	1
NN	\			
NN	n			
VBG	\			
RB	nPoorly			
NN	Diagnosis	NNP	Diagnosis\n\nMantle	2
NN	\			
NN	n			
VBG	\			
JJ	nMantle			
-LRB-	(	NN	(MCL)	2
NN	MCL			
-LRB-	)			
NN	Diagnosis	NNP	Diagnosis\n\nWell	1
NN	\			
NN	n			
VBG	\			
NN	nWell			
NN	Diagnosis	NNP	Diagnosis\n\ninvasive	2
NN	\			
NN	n			
VBG	\			
JJ	ninvasive			
NN	Diagnosis	NNP	Diagnosis\n\nSuperficial	2
NN	\			
NN	n			
VBG	\			
JJ	nSuperficial			
NN	Diagnosis	NNP	Diagnosis	2
JJ	Scanty	NNP	Scanty	1
MD	can	MD	cannot	1
RB	not			
JJ	crypt	NN	crypt	1
NN	Paneth	NNP	Paneth	1
JJ	basal	NN	basal	1
NNS	Granulomas	NNP	Granulomas	1
JJ	histopathologic	NN	histopathologic	2
NN	Diagnosis	NNP	Diagnosis\n\nHyperplastic	2
NN	\			

NN	n			
VBG	\			
JJ	nHyperplastic			
JJ	Hyperplastic	NNP	Hyperplastic	1
NN	Diagnosis	NNP	Diagnosis\n\nTubular	4
NN	\			
NN	n			
VBG	\			
JJ	nTubular			
NN	Diagnosis	NNP	Diagnosis\n\nEndometrail	2
NN	\			
NN	n			
VBG	\			
JJ	nEndometrail			
,	,	VBD	,/endometriod	2
:	/			
JJ	endometriod			
NN	Diagnosis	NNP	Diagnosis\n\nModerately	3
NN	\			
NN	n			
VBG	\			
RB	nModerately			
VBN	differentiated	VBD	differentiated	3
NN	adenocarcinoma	DT	adenocarcinoma	3
NN	adenocarcinoma	VBD	adenocarcinoma	1
NN	Diagnosis	NNP	Diagnosis\n\nFocal	1
NN	\			
NN	n			
VBG	\			
RB	nFocal			
JJ	submucosal	NN	submucosal	1
NN	Comment	NNP	Comment	1
JJ	dysplasia/intramucosal	NN	dysplasia/intramucosal	1
NNS	Fragments	NNP	Fragments	1
NN	Diagnosis	NNP	Diagnosis\n\nColon	1
NN	\			
NN	n			
VBG	\			
NN	nColon			
NN	Rectum	NNP	Rectum	1
NNP	Right	RB	Right	1
JJ	Transverse	NNP	Transverse	1
NN	Sigmoid	NNP	Sigmoid	1