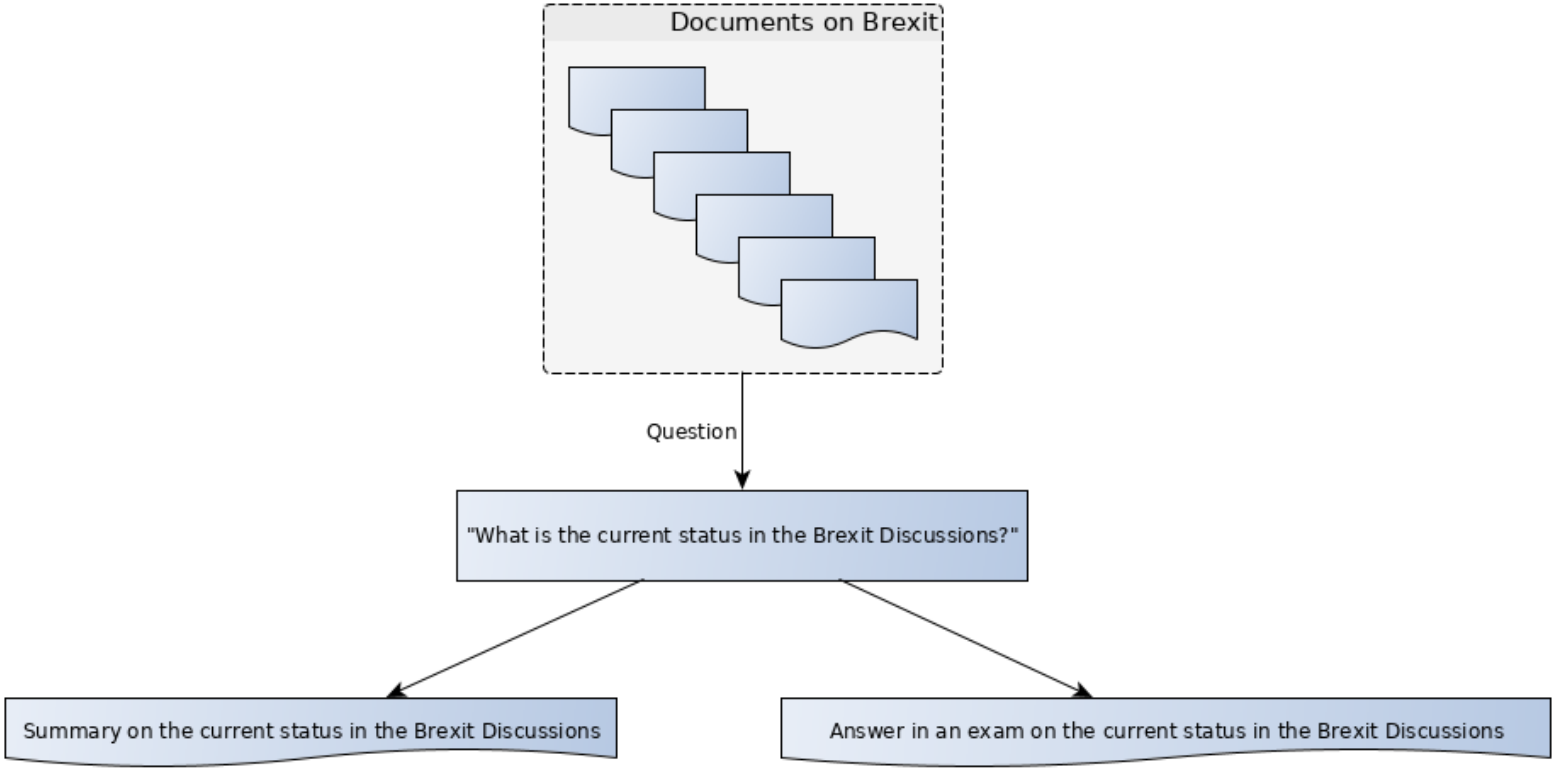
Summarization Evaluation meets Short-Answer Grading

Margot Mieskes and Ulrike Padó

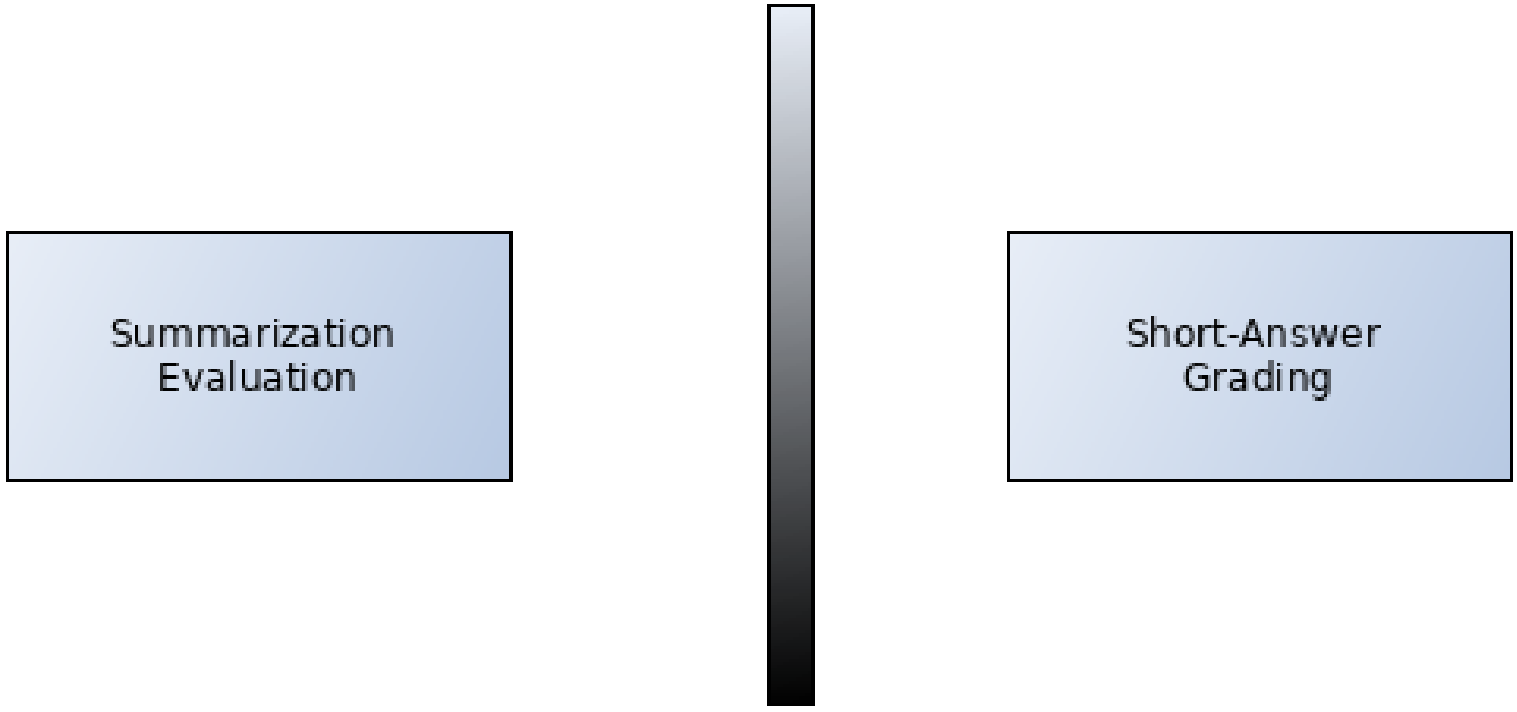
Presentation at NLP4CALL 2019

30 September 2019

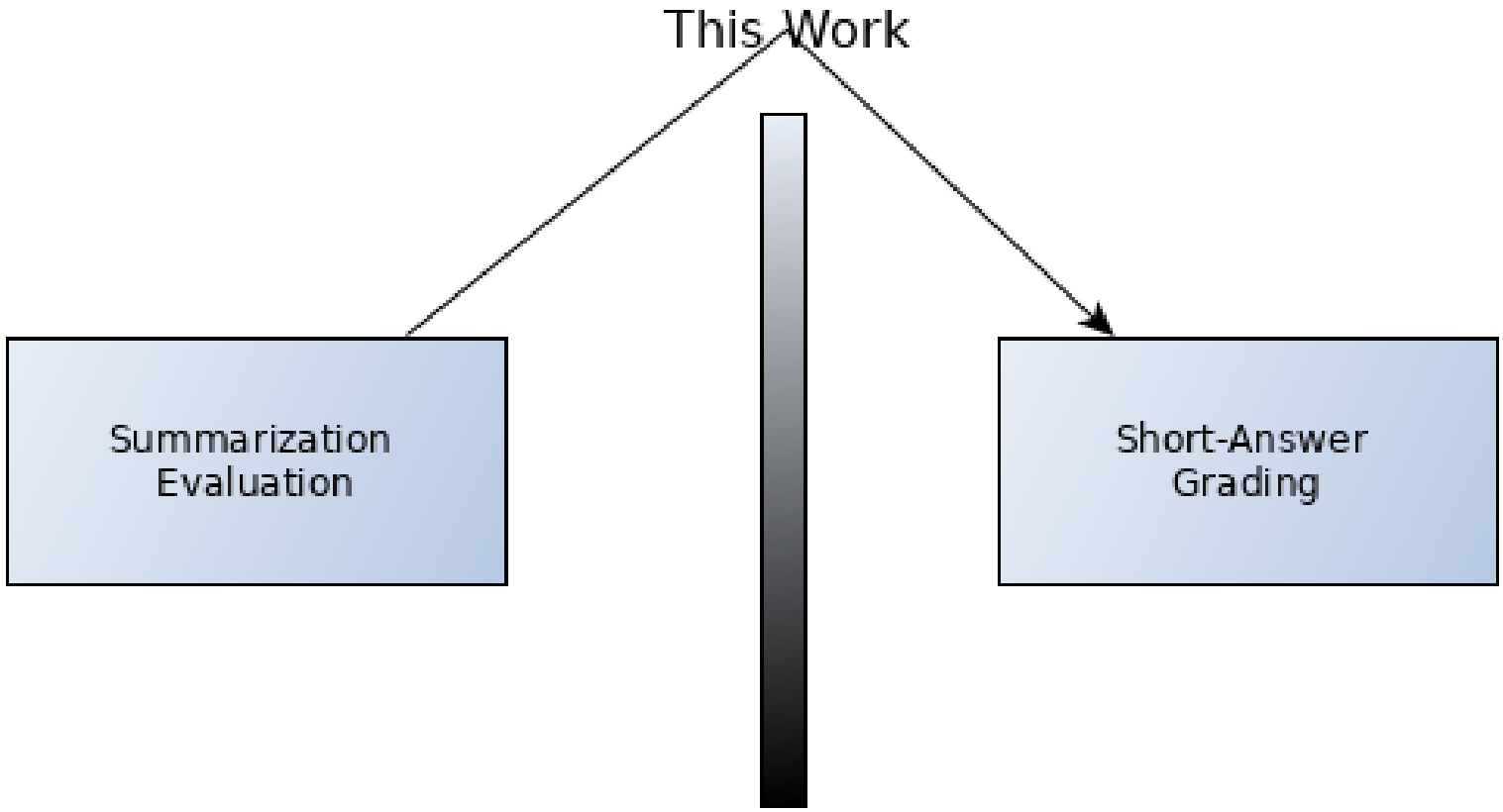
Motivation


Documents on Brexit
Question
"What is the current status in the Brexit Discussions?"
Summary on the current status in the Brexit Discussions
Answer in an exam on the current status in the Brexit Discussions

Motivation



Motivation



Intro to ROUGE

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

ROUGE	Recall Oriented Understudy for Gisting Evaluation
ROUGE-1	Unigram Overlap
ROUGE-2	Bigram Overlap
Stemming	using Wordnet entries
Human Evaluation	Shows high correlation with human evaluation

Intro to ROUGE

X: [A B C D E F G]

Y_1: [A B C D H I K]

Y_2: [A H B K C I D]

Intro to ROUGE

S1. police killed the gunman
S2. police kill the gunman
S3. the gunman kill police
S4. the gunman police killed

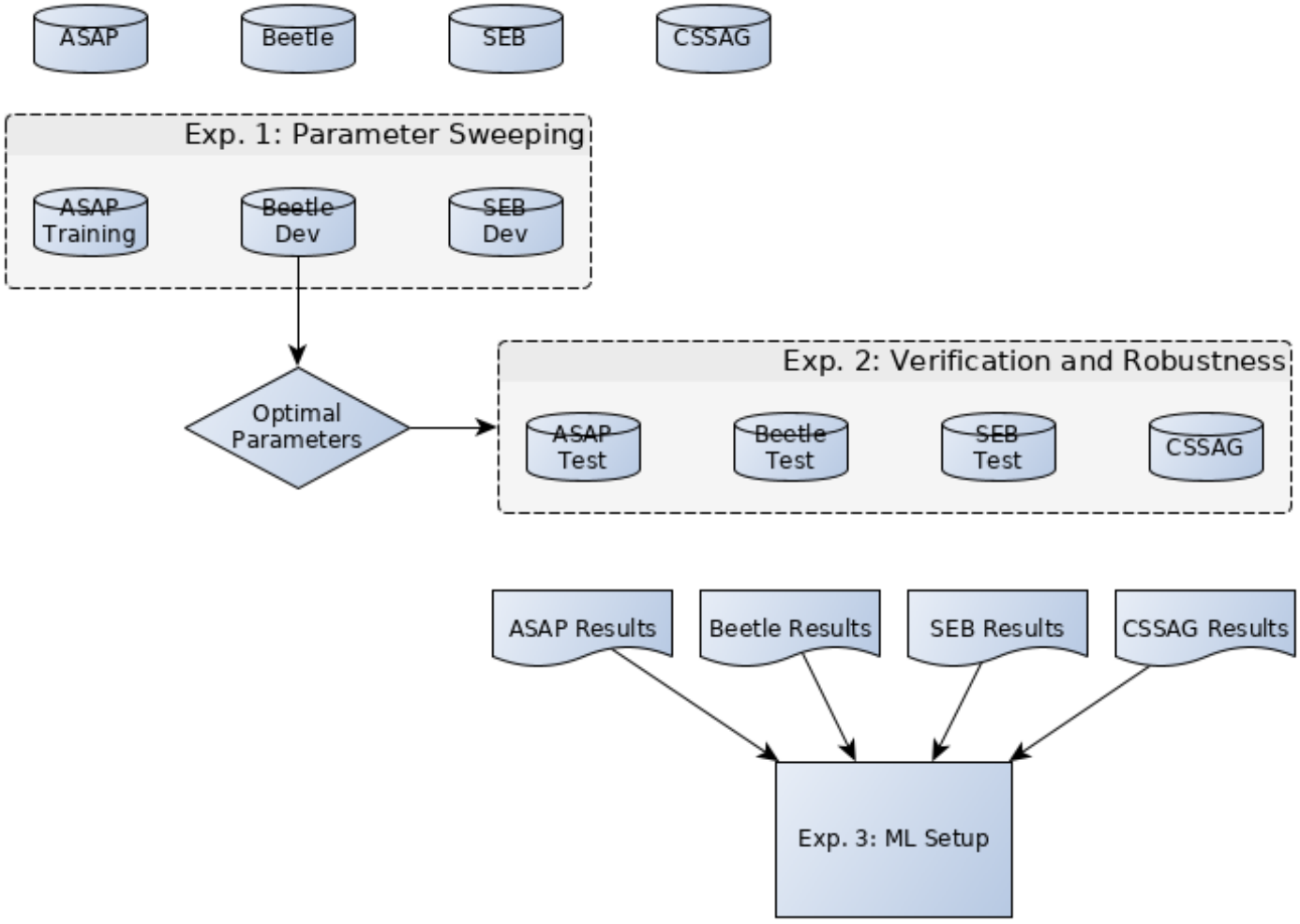
ROUGE Parameters


```
  [-c cf]
  [-d (print per evaluation scores)]
  [-e ROUGE_EVAL_HOME]
  [-h (usage)]
  [-H (detailed usage)]
  [-b n-bytes|-l n-words]
  [-m (use Porter stemmer)]
  [-n max-ngram]
  [-s (remove stopwords)]
  [-r number-of-samples (for resampling)]
  [-2 max-gap-length (if < 0 then no gap length limit)]
  [-3 <H|HM|HMR|HM1|HMR1|HMR2> (for scoring based on BE)]
  [-u (include unigram in skip-bigram) default no)]
  [-U (same as -u but also compute regular skip-bigram)]
  [-w weight (weighting factor for WLCS)]
  [-v (verbose)]
  [-x (do not calculate ROUGE-L)]
  [-f A|B (scoring formula)]
  [-p alpha (0 <= alpha <=1)]
  [-t 0|1|2 (count by token instead of sentence)]
  [-z <SEE|SPL|ISI|SIMPLE>]
  <ROUGE-eval-config-file> [<systemID>]\n
```

ROUGE Setup


```
<ROUGE-EVAL version="1.0">
<EVAL ID="1">
<PEER-ROOT>
/home/mieskes/work/tools/rouge_perl/RELEASE-1.5.5/sample-test/SL2003/systems
</PEER-ROOT>
<MODEL-ROOT>
/home/mieskes/work/tools/rouge_perl/RELEASE-1.5.5/sample-test/SL2003/models
</MODEL-ROOT>
<INPUT-FORMAT TYPE="SEE">
</INPUT-FORMAT>
<PEERS>
<P ID="11">SL.P.10.R.11.SL062003-01.html</P>
<P ID="12">SL.P.10.R.12.SL062003-01.html</P>
<P ID="13">SL.P.10.R.13.SL062003-01.html</P>
<P ID="14">SL.P.10.R.14.SL062003-01.html</P>
<P ID="21">SL.P.10.R.21.SL062003-01.html</P>
<P ID="22">SL.P.10.R.22.SL062003-01.html</P>
<P ID="23">SL.P.10.R.23.SL062003-01.html</P>
<P ID="24">SL.P.10.R.24.SL062003-01.html</P>
</PEERS>
<MODELS>
<M ID="A">SL.P.10.R.A.SL062003-01.html</M>
<M ID="B">SL.P.10.R.B.SL062003-01.html</M>
<M ID="C">SL.P.10.R.C.SL062003-01.html</M>
<M ID="D">SL.P.10.R.D.SL062003-01.html</M>
</MODELS>
</EVAL>
<EVAL ID="2">
<PEER-ROOT>
/home/mieskes/work/tools/rouge_perl/RELEASE-1.5.5/sample-test/SL2003/systems
</PEER-ROOT>
<MODEL-ROOT>
/home/mieskes/work/tools/rouge_perl/RELEASE-1.5.5/sample-test/SL2003/models
</MODEL-ROOT>
<INPUT-FORMAT TYPE="SEE">
</INPUT-FORMAT>
<PEERS>
<P ID="11">SL.P.10.R.11.SL062003-02.html</P>
<P ID="12">SL.P.10.R.12.SL062003-02.html</P>
<P ID="13">SL.P.10.R.13.SL062003-02.html</P>
```

Experimental Setup



Results

Experiment 1: Optimal Parameters

	ASAP	**Beetle**	**SEB**
Stemming	**y**	n	**y**
Stopwords	**n**	**n**	y
ROUGE	**S***	**S***	LCS
Eval Basis	**s**	**s**/t	**s**/t
Model	**best**	all	all
Measure	**R**	$F_{0.5}$	$F_{0.5}$
Conf Level	**95**	**95**/99	**95**/99
optimal τ	0.581	0.469	0.313
final τ	0.581	0.449	0.286

Results

Experiment 2: Parameter Verification & Robustness

Corpus	τ **dev**	τ **test**	**Language**
ASAP	0.581	0.356	
Beetle	0.449	0.306	EN
SEB	0.286	0.223	
CSSAG	–	0.385	DE

Results

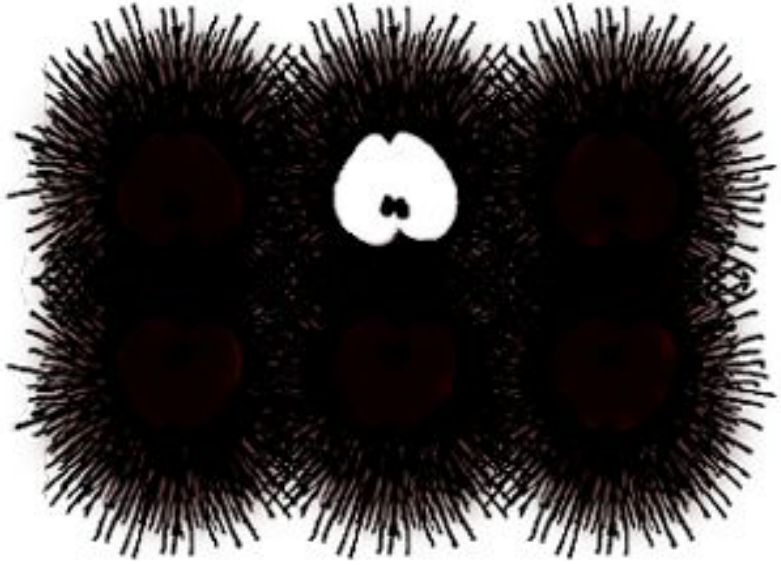
Experiment 3: ML Experiments

	Majority	**Unigram**		**ROUGE**		**NLP**		**NLP+R**	
		RF	SVM	RF	SVM	RF	SVM	RF	SVM
ASAP	58.1	86.3	70.1	80.7	64.0	**86.8**	69.4	86.4	69.9
Beetle	42.6	72.8	71.3	60.7	55.1	**73.6**	73.0	71.9	72.6
SEB	43.7	59.7	65.1	61.4	58.1	66.7	65.2	**67.0**	64.7
CSSAG	45.3	66.2	**70.1**	67.7	64.0	67.6	69.4	68.3	69.9

Discussion & Conclusions

- Summarization and SAG can benefit
- ROUGE can be used in a SAG task
- Parameters used in Summarization are also stable in SAG
- ROUGE can be used for SAG in various languages
- ROUGE results can be used in an ML framework
- ROUGE can serve as a well-defined, reproducible baseline for SAG

Questions, Remarks, Comments, Suggestions?



http://www.userfriendly.org/

Thank you!