COMP4650/6490 Document Analysis

Evaluation in NLP

ANU School of Computing

Administrative matters

- Feedback survey 2
  - Opens: Monday, 16 October 2023, 11:00 AM
  - Closes: Monday, 23 October 2023, 5:00 PM
- Practice exercises
  - Will be released soon
- Drop-in sessions
  - 1pm - 2pm Friday 20 October
  - 1pm - 3pm Thursday 26 October
  - 1pm - 2pm Friday 27 October
  - Room 3.41, Level 3, Hanna Neumann Building #145

Outline

- General evaluation techniques
- Term overlap metrics
- What to compare

Outline

- General evaluation techniques
- Term overlap metrics
- What to compare

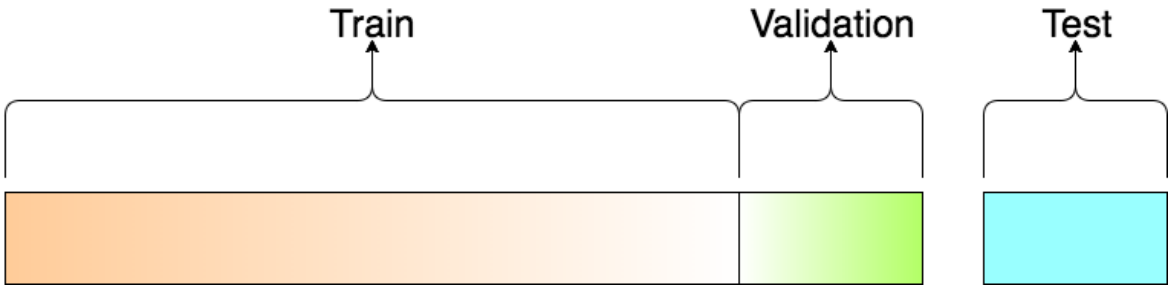
- Intrinsic evaluation
  Directly test a task correctness using a gold standard (also called ground truth), e.g.
  - Evaluate a POS-Tagger
  - Calculate the match between predicted and gold standard POS-tags

- Extrinsic evaluation
  Test whether the output is useful for downstream tasks, e.g.
  - Evaluate a summarisation technique in an IR setting
  - Does using summaries instead of complete documents help a particul ar retrieval task?

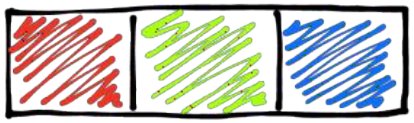
- NLP methods are usually intrinsically evaluated against a gold standard

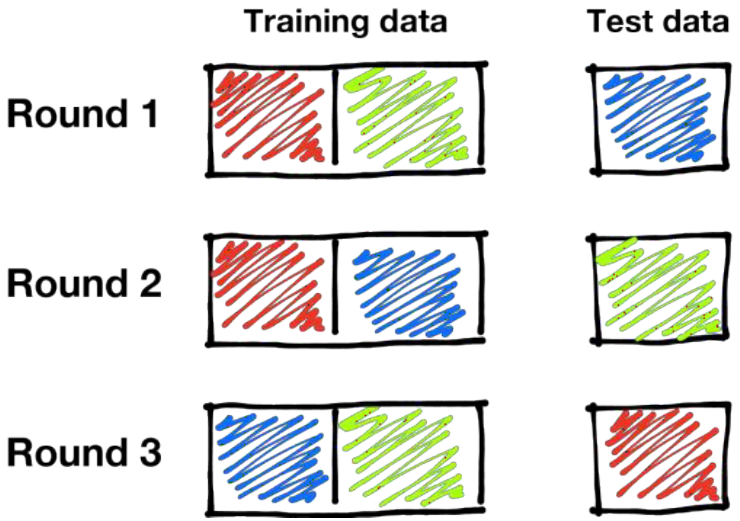
Evaluation methodology recap

- Always split your data in three sets:



- Never test with the training set
- With small data sets, use cross-validation





Classification metrics recap

- Accuracy
  - The number of correct predictions, divided by the total number of instances:

$$\text{accuracy}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{i=1}^{N} \delta(y^{(i)} = \hat{y}^{(i)})$$

  - Not suitable for class-imbalance datasets, e.g.
    - Discover symptoms of a rare disease (positive class), which only appears in 1% of the dataset. If all instances are classified as negative the accuracy is 99%, but it's useless.
- Precision, Recall, F-measure
  - There are two possible errors (FP, FN), and two ways to be correct (TP, TN)
    - False positive (FP): the system incorrectly predicts the label
    - False negative (FN): the system incorrectly fails to predict the label
    - True positive (TP): the system correctly predicts the label
    - True negative (TN): the system correctly predicts that the label does not apply to this instance

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{F}_\alpha = \frac{1}{\frac{\alpha}{\text{Precision}} + \frac{1-\alpha}{\text{Recall}}}, \quad \alpha \in [0,1]$$

Evaluating multi-class classification

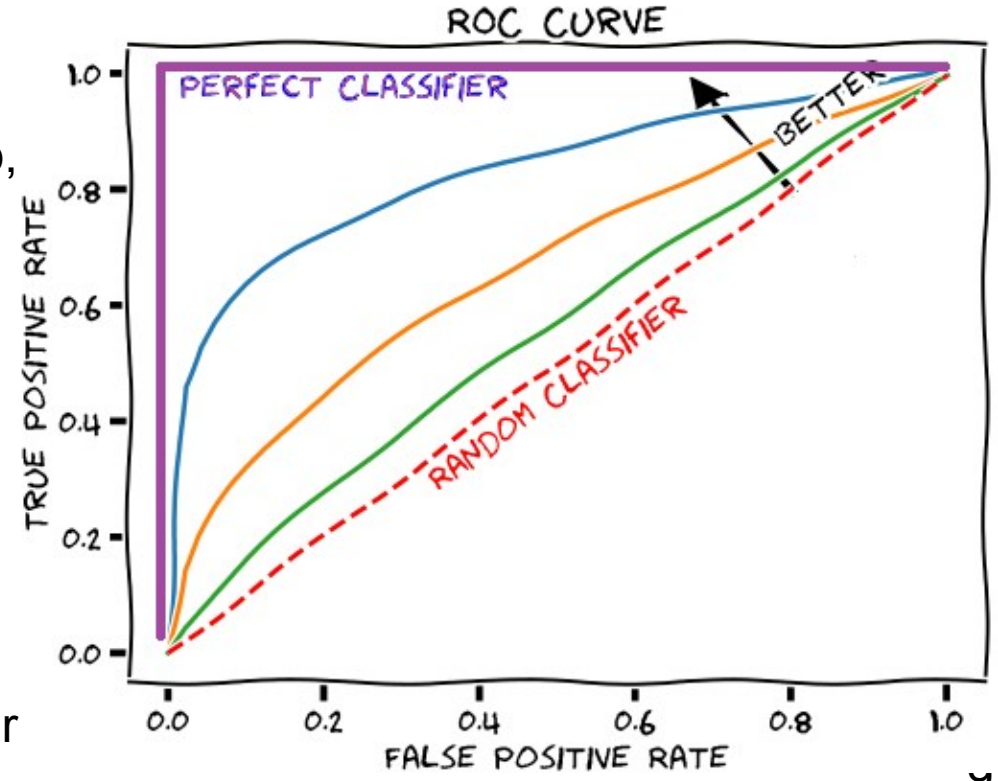
- Macro F-measure:
  When there are multiple labels of interest (e.g. in word sense disambiguation), it is necessary to combine the F-measure across each class.

$$\text{Macro-F}_\alpha(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{K} \sum_{k=1}^{K} \text{F}_\alpha(y_k, \hat{y}_k)$$

- Micro F-measure:
  Counting the total true positives, false negatives and false positives globally, then calculate the F-measure

Threshold free metric: ROC-AUC

- AUC: Area Under The Curve
- ROC: Receiver Operating Characteristics
- For binary classification, it allows to tradeoff between Precision and Recall
  - FPR = FP / (FP + TN)
  - TPR = TP / (TP + FN)
  - AUROC of 0.5 (area under the red dashed line) corresponds to a coin flip, i.e. a random model
  - AUROC less than 0.7 is usually sub-optimal performance
  - AUROC of 0.70 to 0.80 is often good performance
  - AUROC greater than 0.8 is often excellent performance
  - AUROC of 1.0 (area under the purple line) corresponds to a perfect classifier



Examples of evaluation in NLP tasks

- Semantic parsing
  - Accuracy
  - A logical formula is correct or not, an SQL is correct or not. No partially correct cases
- Dependency parsing
  - Precision and Recall of UAS or LAS
  - Unlabelled attachment score (UAS)
    - Evaluates the tree structure
    - Proportion of words whose head is correctly assigned
  - Labelled attachment score (LAS)
    - Evaluates the tree and the arcs relations
    - Proportion of words whose head is correctly assigned with the right dependency label
- Named entity recognition
  - Macro Precision/Recall (as with many other multi-class classification problems)
- Ranking (e.g. Coreference resolution)
  - Mean Average Precision
  - Mean Reciprocal Rank

Australian National University

Outline

- General evaluation techniques
- Term overlap metrics
- What to compare

When the output is a text sequence

The output could be a text sequence

- Machine translation
- Text generation
- Text simplification
- Image Caption Generation

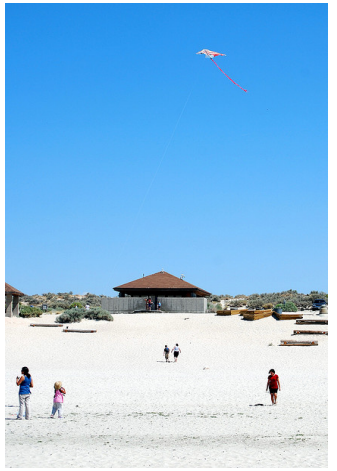
Challenges:

- Almost infinite possible correct outputs
- Unlikely that the model will give the same result as a human labeller
- Getting people to manually evaluate every output is costly

When the output is a text sequence

- Consider image captioning where the task is to generate a caption that describes an image:
  - The input is an image, the output is a sentence
  - Each image in the test set has multiple human written captions
  - We want to compare a generated caption to see how closely it matches the references

Ref1: A cat laying on a couch beside a remote.
Ref2: A cat sitting on the couch.
Ref3: A cat lounging on the couch next to a remote.



Ref1: Children flying a kite at the beach.
Ref2: Some children with a kite at the beach.
Ref3: A kite in the sky with some people beneath.

- The same sort of evaluation setting is used in tasks such as machine translation. (Though the input is a sequence of text rather than an image)
- Note: we don't always have multiple references, sometimes only one is available.

Term overlap evaluation metrics

Use term overlap metrics to evaluate:

- Machine translation
- Text summarisation
- Text simplification
- Q & A
- Chatbots
- Caption generation
- …

Popular term overlap metrics:

- BLEU
- ROUGE
- …

Term overlap evaluation metrics

BLEU (bilingual evaluation understudy)

- A modified form of precision to compare a candidate translation against multiple reference translations.
- Perfect match = 1.0, Perfect mismatch = 0.0
- Precision-based metrics are biased in favour of short translations → brevity penalty (BP)

$$p_n = \frac{\text{number of } n\text{-grams appearing in both reference and hypothesis translations}}{\text{number of } n\text{-grams appearing in the hypothesis translation}}.$$

- BLEU is based on $\exp \frac{1}{N} \sum_{n=1}^{N} \log p_n$ (usually with smoothing to avoid log(0))

Example:

- A reference translation and three system outputs
- For each output, p_n indicates the precision at each n-gram, BP indicates the brevity penalty.

	Translation	p_1	p_2	p_3	p_4	BP	BLEU
Reference	*Vinay likes programming in Python*						
Sys1	*To Vinay it like to program Python*	$\frac{2}{7}$	0	0	0	1	.0
Sys2	*Vinay likes Python*	$\frac{3}{3}$	$\frac{1}{2}$	0	0	.51	.0
Sys3	*Vinay likes programming in his pajamas*	$\frac{4}{6}$	$\frac{3}{5}$	$\frac{2}{4}$	$\frac{1}{3}$	1	.51

(no smoothing)

Term overlap evaluation metrics

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation)
  - Calculate the recall between human and automatic outputs in terms of n-grams (n-gram overlap)
  - ROUGE-N: Overlap of n-grams between the system and a gold standard reference
  - Often, we use both BLEU and ROUGE.
  - ROUGE is used instead of BLEU in summarisation and text simplification tasks because getting high precision is easy (e.g. in summarisation just copy the first sentence verbatim)

Australian National University

Outline

- General evaluation techniques
- Term overlap metrics
- What to compare

Classifier comparison

- Comparison between algorithms, e.g.
  - Logistic regressions vs. MLP
  - L2 regularisation vs. L1 regularisation
- Comparison between feature sets, e.g.
  - Bag-of-words vs. Word embeddings
  - Word embeddings vs. Character embeddings

Baselines

- How can we know if our result is good?
- Evaluate against a set of baselines:
  - trivial models (e.g. always predict most common class, random guessing)
  - simple models (e.g. logistic regression)
  - well known methods for your task
  - the current state of the art method

	BLEU-4	ROUGE
(1) Biten (Avg + CtxIns) [3]	0.89	12.2
(2) Biten (TBB + AttIns) [3]	0.76	12.2
(3) LSTM + GloVe + IA	1.97	13.6
(4) Transformer + GloVe + IA	3.48	17.0
(5) LSTM + RoBERTa + IA	3.45	17.0

Use multiple different evaluation metrics.

Example from (Tran et al. 2020) https://arxiv.org/abs/2004.08070

Ablation studies

- How do we prove that all parts of the proposed model are necessary?
- Ablation testing involves systematically removing (ablating) various aspects of a model, such as feature groups, to see if the ablated classifier is as good as the full model

	BLEU-4	ROUGE
(6) Transformer + RoBERTa	4.60	18.6
(7) + image attention	5.45	20.7
(8) + weighted RoBERTa	6.0	21.2
(9) + face attention	**6.05**	**21.4**
(10) + object attention	**6.05**	**21.4**

Example from (Tran et al. 2020) https://arxiv.org/abs/2004.08070

Different datasets

- How do we prove that our model is generally applicable to a problem not just a single dataset?
- Test on multiple different datasets
- Make sure to include any standard benchmark datasets

	NER				
	de	de_{06}	en	es	nl
Baevski et al. (2019)	-	-	93.5	-	-
Straková et al. (2019)	85.1	-	93.4	88.8	92.7
Yu et al. (2020)	86.4	90.3	93.5	90.3	93.7
Yamada et al. (2020)	-	-	94.3	-	-
XLM-R+Fine-tune	87.7	91.4	94.1	89.3	95.3
ACE+Fine-tune	**88.3**	**91.7**	**94.6**	**95.9**	**95.7**

Example from (Wang et al. 2020) https://arxiv.org/pdf/2010.05006.pdf

Summary

- General evaluation techniques
- Term overlap metrics
- What to compare

References

- Chapter 13. Speech and Language Processing (3rd ed. draft)
- Chapter 18. Natural language processing (by Jacob Eisenstein). 2018.
- Wang, Xinyu, et al. "Automated Concatenation of Embeddings for Structured Prediction." Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021.
- Tran, Alasdair, Alexander Mathews, and Lexing Xie. "Transform and tell: Entity-aware news image captioning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.