Foundations of Natural Language Processing

Lecture 3: Corpora and Experimental Design

Mirella Lapata
School of Informatics
University of Edinburgh
mlap@inf.ed.ac.uk

Slides based on content from: Philipp Koehn, Alex Lascarides, Sharon Goldwater, Shay Cohen, Khalil Sima'an, Ivan Titov

Corpora in NLP

This lecture:

- What is a corpus?
- Why do we need **text corpora** for NLP? (learning, evaluation)
- What is **experimental design** in NLP?
- What are the principles behind **model evaluation**?

Corpora in NLP

corpus: noun, plural corpora or, sometimes, corpuses.

1. a large or complete collection of writings: the entire corpus of Old English poetry.
2. the body of a person or animal, especially when dead.
3. Anatomy. a body, mass, or part having a special character or function.
4. Linguistics. a body of utterances, as words or sentences, assumed to be representative of and used for lexical, grammatical, or other linguistic analysis.
5. a principal or capital sum, as opposed to interest or income.

Dictionary.com

Corpora in NLP

- To understand and model how language works, we need empirical evidence. Ideally, naturally-occurring corpora serve as realistic samples of a language.
- Aside from linguistic utterances, corpora include **metadata**: side information about where the language comes from, such as author, date, topic, publication.
- Of interest for NLP are corpora with **linguistic annotations**: where humans have read the text and marked categories or structures describing their syntax and/or meaning, or right answer.

Elements of Corpus Design

- **Text Sampling:** make sure that the corpus reflects the appropriate *language diversity*, choose a representative and systematic *selection technique*. Think about whether texts will be chosen at random, on purpose, or through stratified sampling.
- **Corpus Size and Balance:** determine the appropriate *corpus size* while considering computational capabilities and research objectives. Make sure the corpus has a diverse range of language attributes, including rare or uncommon events.
- **Text Annotation:** choose the appropriate *level of linguistic annotation*, which may involve part-of-speech tagging, named entity recognition, parse trees, sentiment analysis, or semantic annotation. Decide whether semi-automatic, or manual annotation will be used.

Examples of Corpora

- **BookCorpus:** 7,000 self-published books, 985 million words.

- **Brown:** 1M words in 15 genres. POS-tagged. **SemCor** subset (234K words) labeled with WordNet senses.
- **WSJ:** 6 years of Wall Street Journal; used to create Penn Treebank, PropBank, and more!
- **BNC**: 100M words; balanced selection of written and spoken genres.
  Gigaword: 1B words of news text.
- **Common Crawl:** since 2008, created by crawling the Internet (petabytes of data).
- **Wikipedia:** as of 16 October 2024, 24.09 GB compressed without media.
- **OpenSubtitles:** subtitles from movies and TV shows, 7.2 GB of data.

Annotation

Suppose you are tasked with building an annotated corpus (e.g., with part-of-speech tags) In order to estimate cost in time and money, you need to decide on:

- Source data (genre? size? licensing?)
- Annotation scheme (complexity? guidelines?)
- Annotators (expertise? training?)
- Annotation software (graphical interface?)
- Quality control procedures (multiple annotation, adjudication?)

Assuming a competent annotator, some kinds of annotation are straightforward, while some are not (ambiguity, gray areas between categories in the annotation scheme).

You play annotator

Verb, noun, or adjective?

- We had been **walking** quite briskly.
- **Walking** was the remedy, they decided.
- In due time Sandburg was a **walking** thesaurus of American folk music.
- We all lived within **walking** distance of the studio.
- A woman came along carrying a folded umbrella as a **walking** stick.
- The **Walking** Dead premiered in the U.S. on October 31, 2010, on the cable television channel AMC.

Annotation Guidelines

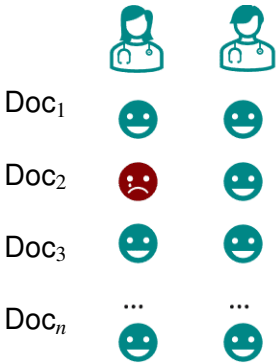
- **Penn Treebank:** 36 POS tags (excluding punctuation).
- Tagging guidelines (3rd Revision): 34 pages

  > *The temporal expressions yesterday, today and tomorrow should be tagged as nouns (NN) rather than as adverbs (RB). Note that you can (marginally) pluralize them and that they allow a possessive form, both of which true adverbs do not. (p. 19)*

- An entire page on nouns vs. verbs.
- 3 pages on adjectives vs. verbs.
- Penn Treebank bracketing (tree) guidelines: >300 pages!

Even with extensive guidelines, human annotations won't be perfect: simple error (hitting the wrong button), not reading the full context, forgetting a detail from the guidelines, cases not anticipated by or not fully specified in guidelines.

Inter-annotator Agreement



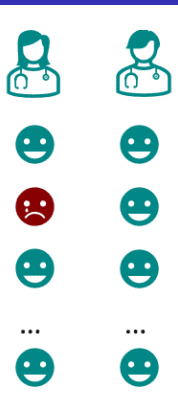
Inter-annotator Agreement

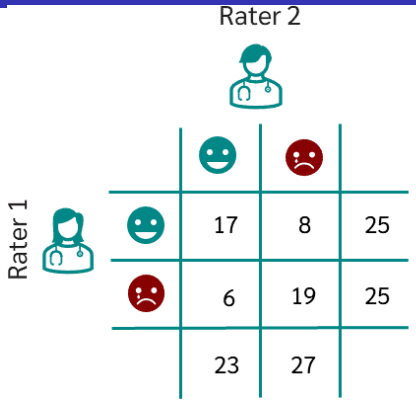
- **Raw agreement rate:** proportion of labels in agreement ($\frac{17+19}{50} = 72\%$)

Inter-annotator Agreement

- **Raw agreement rate:** proportion of labels in agreement ($\frac{17+19}{50} = 72\%$)
- What if some decisions are more frequent than others and raters agree by accident?

Inter-annotator Agreement





$$p_e = \underbrace{\frac{25}{50} \cdot \frac{23}{50}}_{0.5 \cdot 0.46} + \underbrace{\frac{25}{50} \cdot \frac{27}{50}}_{0.5 \cdot 0.54}$$

$$0.23 + 0.27 = 0.5$$

- **Raw agreement rate:** proportion of labels in agreement ($\frac{17+19}{50} = 72\%$)
- What if some decisions are more frequent than others and raters agree by accident?
- **Cohen's Kappa** corrects agreement by hypothetical probability of random match.

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{0.72 - 0.5}{1 - 0.5} = 0.44$$

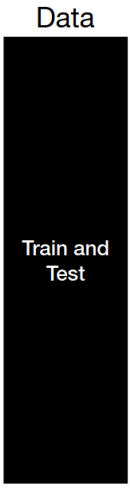
Training with Data

- What is our goal when we train a model?
- We want a model that will preform as good as possible when given **data in the wild**.
- How can we get close to this with the **data we have**?

Training with Data

- What is our goal when we train a model?
- We want a model that will preform as good as possible when given **data in the wild**.
- How can we get close to this with the **data we have**?

1. Train on **all the data and test on all the data** (bad idea, no generalization).



Training with Data

- What is our goal when we train a model?
- We want a model that will preform as good as possible when given **data in the wild**.
- How can we get close to this with the **data we have**?

1. Train on **all the data and test on all the data** (bad idea, no generalization).
2. **Separate train and test** (as we use test data more and more we overfit to it).

Training with Data

- What is our goal when we train a model?
- We want a model that will preform as good as possible when given **data in the wild**.
- How can we get close to this with the **data we have**?

1. Train on **all the data and test on all the data** (bad idea, no generalization).
2. **Separate train and test** (as we use test data more and more we overfit to it).
3. **Development test** distinguishes development testing from real testing.

Training with Data

- What is our goal when we train a model?
- We want a model that will preform as good as possible when given **data in the wild**.
- How can we get close to this with the **data we have**?

1. Train on **all the data and test on all the data** (bad idea, no generalization).
2. **Separate train and test** (as we use test data more and more we overfit to it).
3. **Development test** distinguishes development testing from real testing.
4. **Validation set** can be used for model selection

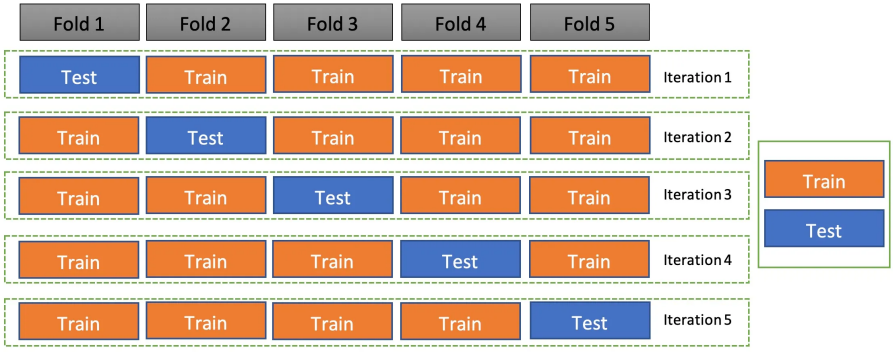
Training with Data

- What is our goal when we train a model?
- We want a model that will preform as good as possible when given **data in the wild**.
- How can we get close to this with the **data we have**?

1. Train on **all the data and test on all the data** (bad idea, no generalization).
2. **Separate train and test** (as we use test data more and more we overfit to it).
3. **Development test** distinguishes development testing from real testing.
4. **Validation set** can be used for model selection
5. Shuffle dev and train once in a while; **touch test as little as possible**.

Cross-validation

What if my dataset is too small to have a nice train/test or train/dev/test split?



- Partition the data into k pieces and treat them as mini held-out sets.
- Each fold is an experiment with different held-out set,
- After k folds, every data point will have a held-out prediction!
- **Still important to have a separate blind test set**. How to choose k (typically 5–10)?

Measuring a Model's Performance

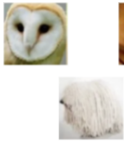
Animal	 Barn owl  Chihuahua  Sheep dog
Not animal	 Apple  Muffin  Mop

Measuring a Model's Performance

Animal	 Barn owl  Chihuahua  Sheep dog
Not animal	 Apple  Muffin  Mop

		Predicted: Animal	Predicted: Not animal
Actual	Animal	True Positives	False Negatives
	Not animal	False Positives	True Negatives

Measuring a Model's Performance

Actual \ Predicted	Animal	Not animal
Animal		
Not animal		

True Positives	3
True Negatives	3
False Positives	0
False Negatives	0

Measuring a Model's Performance

Actual \ Predicted	Animal	Not animal
Animal	 	
Not animal		 

True Positives	2
True Negatives	2
False Positives	1
False Negatives	1

Measuring a Model's Performance

Accuracy: Out all the predictions we made, how many were true?

$$accuracy = \frac{true\ positives + true\ negatives}{true\ positives + true\ negatives + false\ negatives + false\ positives}$$

Precision: Out of all the positive predictions we made, how many were true?

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

Recall: Out of all the data points that should be predicted as true, how many did we correctly predict as true?

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

Measuring a Model's Performance

Precision: Out of all the positive predictions we made, how many were true?

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

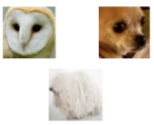
Recall: Out of all the data points that should be predicted as true, how many did we correctly predict as true?

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

F1 Score: combines recall and precision. F1 can therefore be used to measure how effectively our models trade-off precision against recall.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

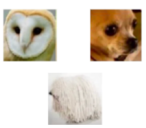
Model 1 (Classifies all images as animal)

		Predicted: Animal	Predicted: Not animal
Actual	Animal		
	Not animal		

True Positives	3
True Negatives	0
False Positives	3
False Negatives	0

Accuracy	50%	$\frac{3+0}{3+0+0+3}$
Precision	50%	$\frac{3}{3+3}$
Recall	100%	$\frac{3}{3+0}$
F1 score	67%	$2 \cdot \frac{0.5 \cdot 1}{0.5+1}$

Model 2 (Classifies all images as not animal)

		Predicted: Animal	Predicted: Not animal
Actual	Animal		
	Not animal		

True Positives	0
True Negatives	3
False Positives	0
False Negatives	3

Accuracy	50%	$\frac{0+3}{0+3+3+0}$
Precision	0%	$\frac{0}{0+0}$
Recall	0%	$\frac{0}{0+3}$
F1 score	0%	

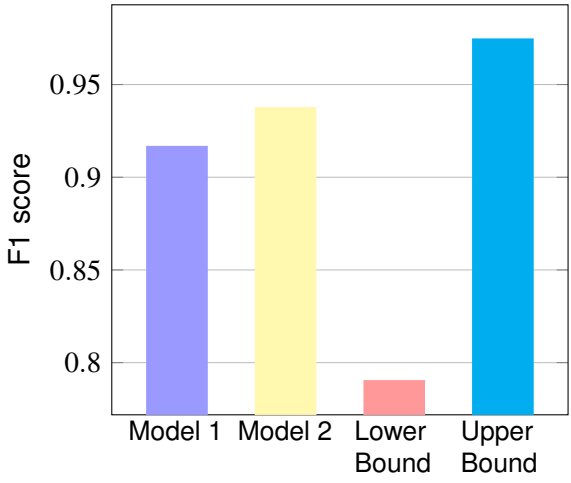
Model 3 (Overpredicts images as not animal)

Actual \ Predicted	Animal	Not animal
Animal	 	
Not animal		  

True Positives	2
True Negatives	3
False Positives	0
False Negatives	1

Accuracy	83%	$\frac{2+3}{2+3+1+0}$
Precision	100%	$\frac{2}{2+0}$
Recall	67%	$\frac{2}{2+1}$
F1 score	80%	$2 \cdot \frac{1 \cdot 0.67}{1+0.67}$

Lower Bound, Upper Bound, and Statistical Significance



- **Lower Bound:** performance of a 'simpler' model (**baseline**) – Model always picks most frequent class (majority baseline).
- **Upper Bound:** When using a human gold standard, check the agreement of humans against that standard
- **Statistical Significance:** Is the difference between Model 1 and Model 2 significant? Are they significantly better than the baseline?

Which Significance Test?

Parametric tests assume that the data approximately follows a normal distribution

- t-test, z-test, ANOVA, ...
- You don't need to know the mathematical formulae; available in statistical libraries!

Non-Parametric tests do not assume anything about the distribution followed by the data

- We usually need non-parametric tests
- Can use Wilcoxon Signed Rank test, McNemar's test or variants of it.
- Stochastic / permutation tests are a convenient alternative (esp. with complex predictions, such as parse trees)

See "Predicting Linguistic Structure", Smith (2011, Appendix B) for a detailed discussion of significance testing methods for NLP.

Take-home message

- NLP models are trained and evaluated on corpora which can have annotations provided by humans following explicit guidelines.
- Inter-annotator agreement measures whether raters can reliably apply annotation guidelines (and also tells us whether the task is feasible).
- Models are trained and tested on different data splits.
- Basic metrics of model performance: accuracy, precision, recall, F1.
- You compare performance of your model against: upper bound, baseline model, someone else's model, and use an appropriate significance test to see if differences are 'real' or within margin of error (i.e., likely due to chance).

Next lecture: we discuss how to build a text classifier.