Natural Language Processing
自然語言處理

黃瀚萱
Department of Computer Science
National Chengchi University
2020 Fall

Lesson 5
Evaluation &
Word Sense Disambiguation

Schedule

Date	Topic
9/16	Introduction
9/23	Linguistic Essentials
9/30	Collocation
10/7	Language Model
10/14	**Word Sense Disambiguation**
10/21	Text Classification (**HW1 will be assigned**)
10/28	Invited Talk: NLP and Cybersecurity (**Term Project**)
11/4	POS Tagging
11/11	Midterm Exam

Schedule

Date	Topic
11/18	Chinese Word Segmentation
11/25	Word Embeddings
12/2	Neural Networks for NLP
12/9	Parsing
12/16	Discourse Analysis
12/23	Invited Talk
12/30	Final Project Presentation I
1/6	Final Project Presentation II
1/13	Final Exam

Agenda

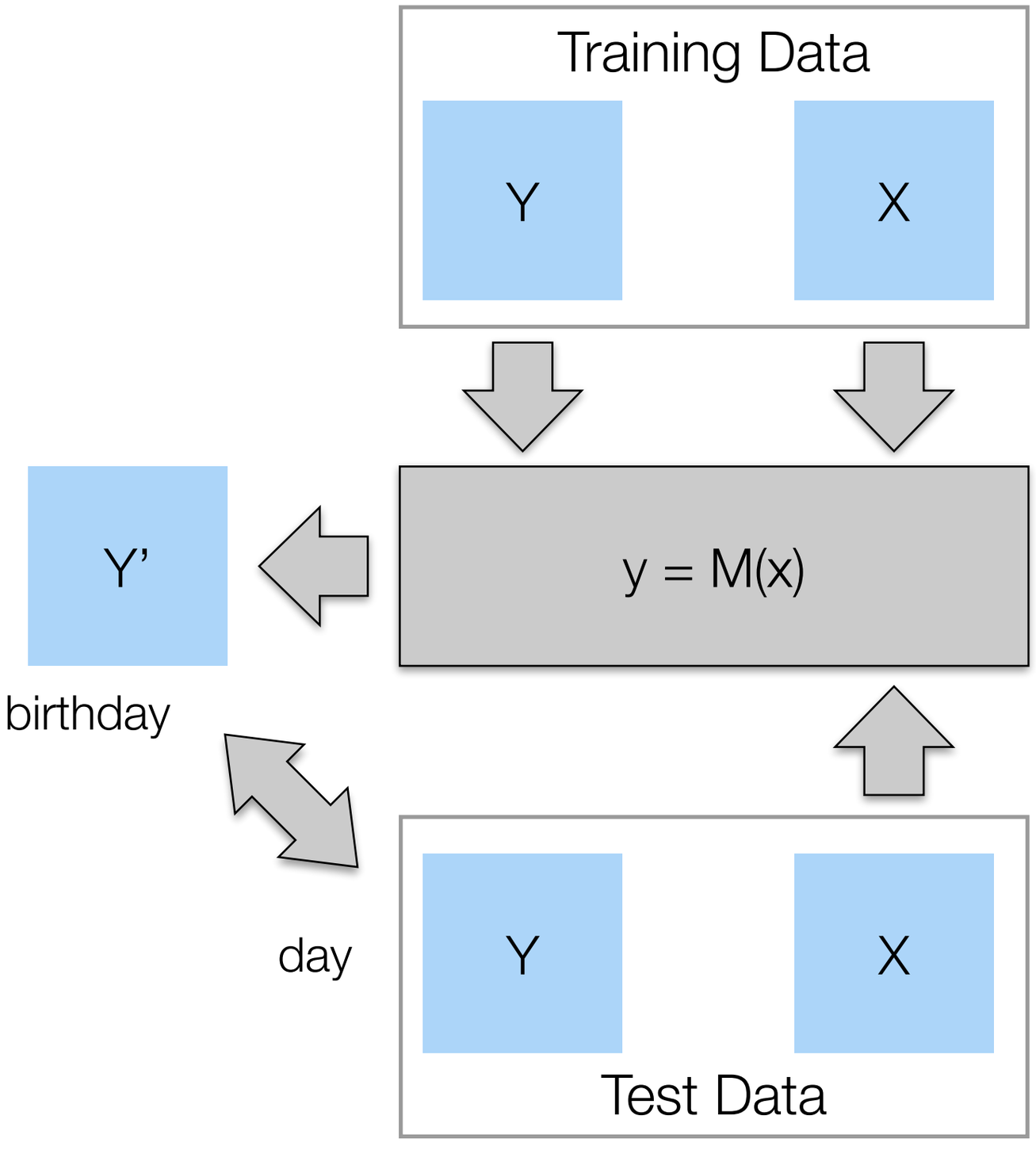
- Performance evaluation
  - Metrics
  - Significant test
- Word sense disambiguation
- Naive Bayes classifier

Performance Evaluation

Performance Evaluation

- Fair test
  - The test data has not been seen before for the model
- Held out data
  - Training data (for parameter optimizing)
  - Validation data (for hyperparameter tuning)
  - Test data (reserved for the final test)

Evaluation with Held Out Data



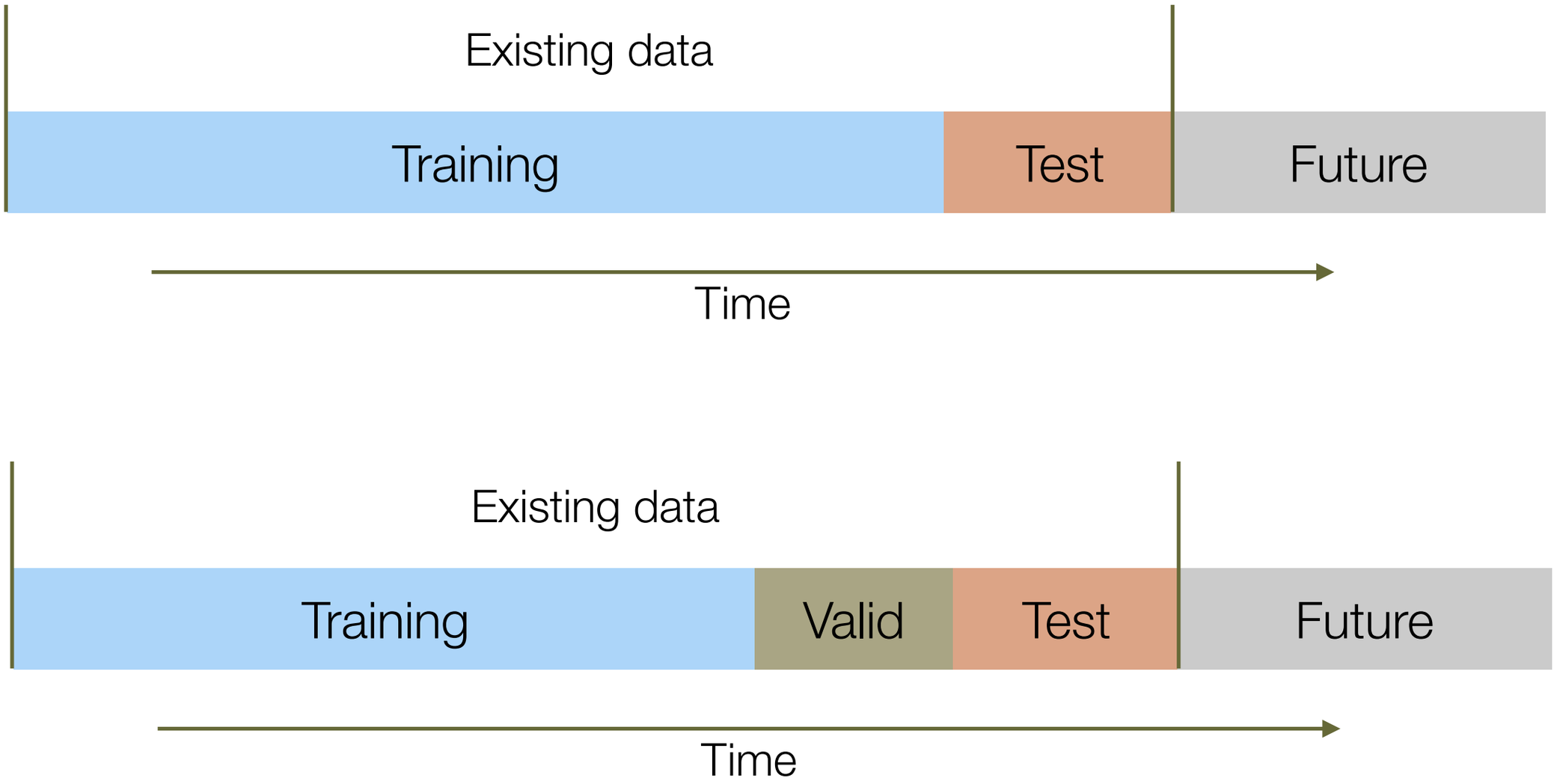
X	Y
My dog ate my	cake
My dog ate my	homework
My father is my	hero
classroom is the thrid	teacher
…	…

X	Y
Today is my	day

Unseen to M

Training Data vs Test Data

- Data split



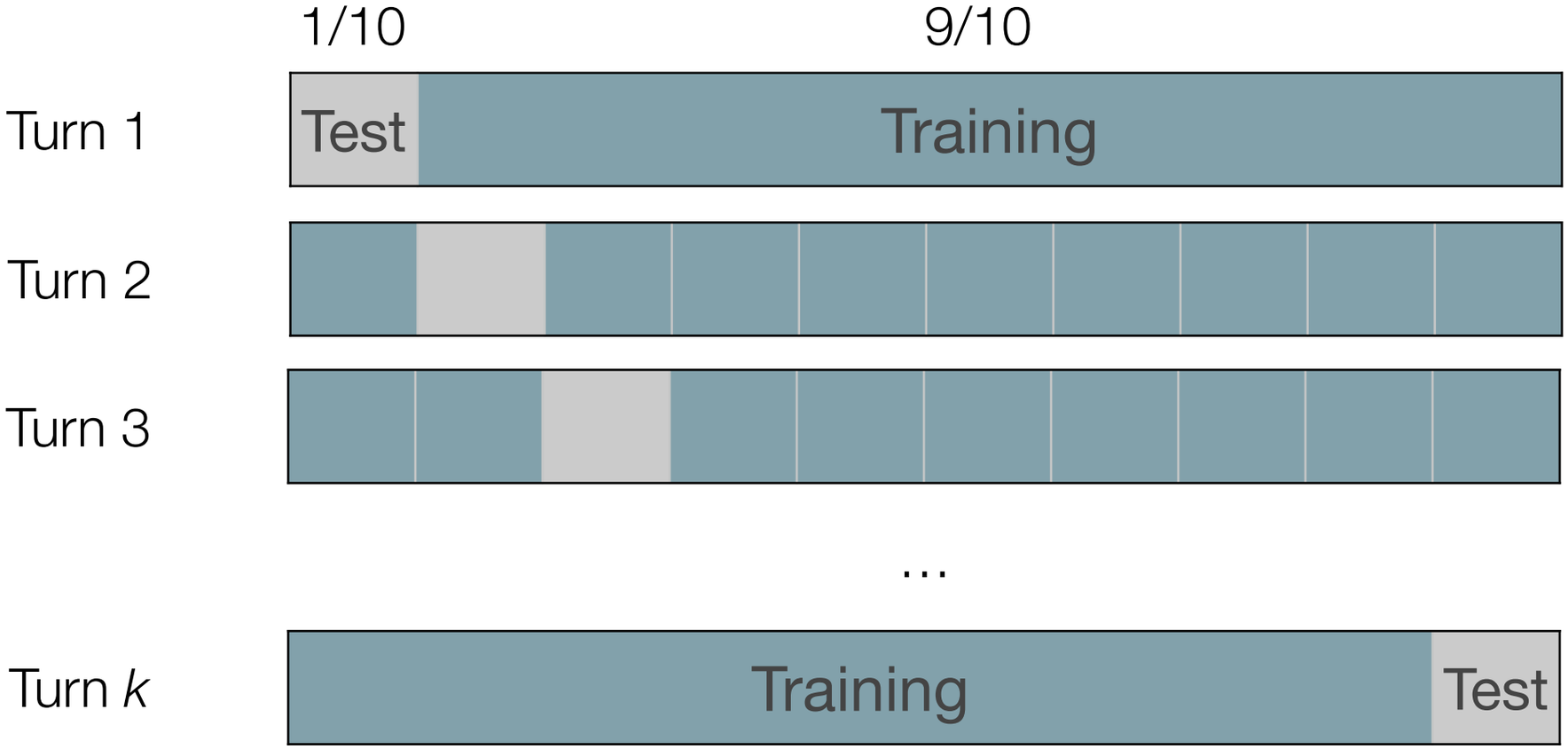
Cross-Validation

- To prevent a relatively large part of the full training set is wasted.

- Two-fold cross validation
  - Split the data into two parts.
  - Two parts take turns as test data.

k-fold Cross-Validation

- Exploiting more training data
- Splitting *n* training instances into *k* folds
  - In each turn, 1/k are used for testing, (k-1) / k for training.

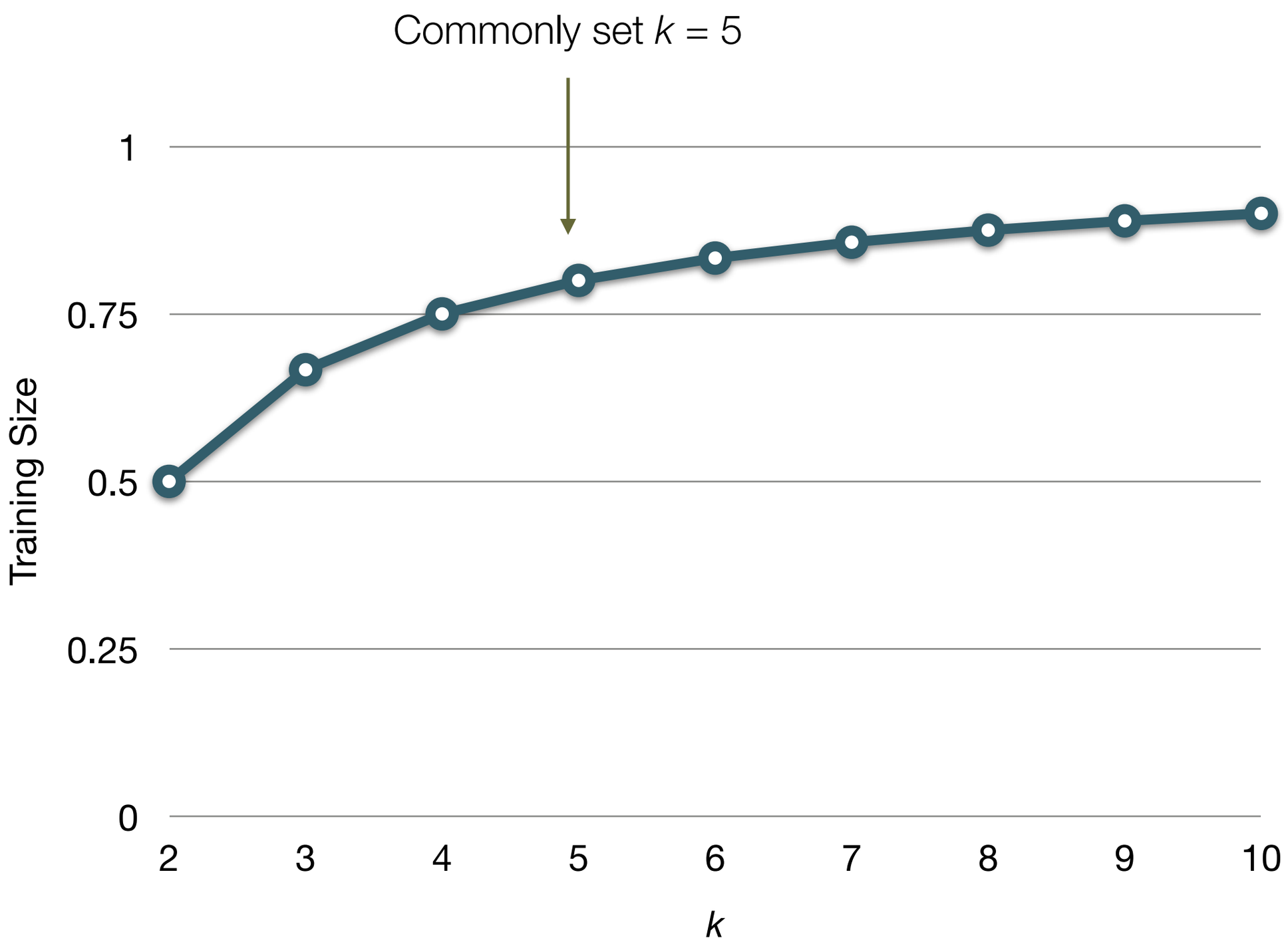


All data will be tested once!

k-fold Cross-Validation

- The larger the *k*
  - The more the training data can be used
    - Marginal effect: 1/2 vs 4/5 vs 9/10
- Time consuming
  - You have train the model *k* times

Marginal Effect with a Larger *k*



Leave-One-Out Cross-Validation

- Leave-One-Out
  - $k = n$, where n is the size of data
  - Takes a long time with a large n
- In practical, the reasonable k is 2, 5, and 10.
  - Five-fold cross-validation is the most popular.

Performance Metrics

- Measuring how good a model is
- Accuracy (正確率)
    - How many times the model makes correct decisions
- Recall (召回率)
    - How many targets the model successfully detects
- Precision (準確率)
    - How confident the model is
- F-score
    - Harmonic average (調和平均數) of Recall and Precision

Accuracy

- The basic and the most fundamental metrics

$$Accuracy = \frac{\text{Correct Predictions}}{\text{All Predictions}}$$

- A bad model may achieve a very high accuracy in some cases.

Problem of Accuracy

- Suppose we are evaluating a model for detecting whether an inbound passenger is a case of COVID-19 at the airport.
    - Only 2% of passengers actually infected with COVID-19
- A model always predicts Negative (without COVID-19 infected) will achieve an accuracy of 98%
    - Even this model is totally useless.
- Data imbalance

假設入境普篩並居家檢疫

2020/08/22

COVID-19(武漢肺炎)

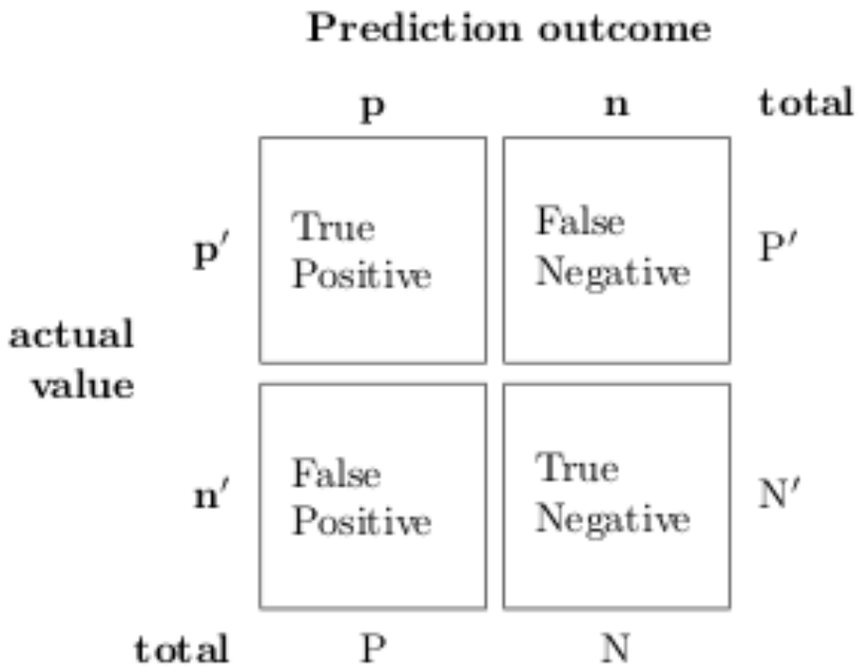
		+	-	
入境普篩 敏感性 90% 特異性 95%	+	450 陽性個案	12,475 偽陽性	12,925
	-	50 偽陰性	237,025 眞正陰性	237,075
		500	249,500	250,000 （單位：人）

若所有入境者皆於機場進行篩檢，檢驗陰性者居家檢疫，陽性者全數收治負壓隔離病房，則可能有12,475名偽陽性明明沒有生病，卻因為檢驗誤差，造成醫療人力與資源的浪費。

中央流行疫情指揮中心

Four Cases of Predictions

- **True-Positive**: the model says Positive (陽性), and the passenger is actually inflected with COVID-19.
- **True-Negative**: the model says Negative (陰性), and the passenger is not inflected with COVID-19.
- **False-Positive** (偽陽性): the model says Positive, but the passenger is actually negative (Waste of the test).
- **False-Negative** (偽陰性): the model says Negative, but the passenger is actually positive (**dangerous!**)

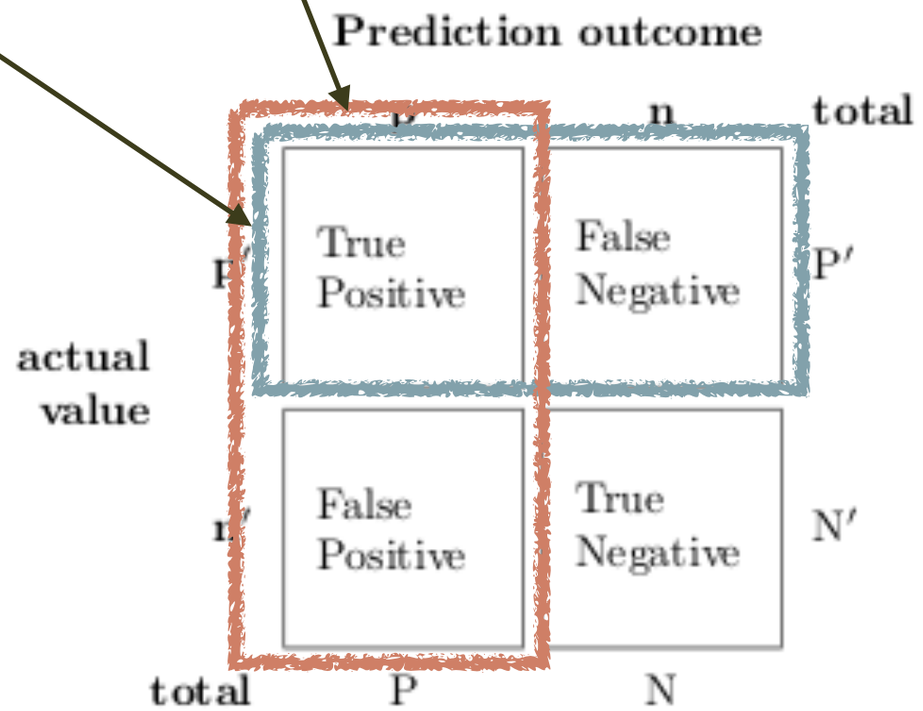


Performance Metrics

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} = \frac{\text{correct}}{\text{total}}$$

$$Precision = \frac{tp}{tp + fp} = \frac{tp}{\text{Predicted as positive}}$$

$$Recall = \frac{tp}{tp + fn} = \frac{tp}{\text{Actually positive}}$$



Precision vs Recall

- A model that always says *Positive* achieves a Recall of 100%.
  - All passengers' bags have to be screened
  - Very disturbing and annoying
- A model that says Positive only when it is most sure will achieve a very high Precision score.
  - Only a few pork can be detected.
- A good model will achieve both high precision and high recall.

F-Score or F1-Measure

- The single indicator that takes both Precision and Recall into account.

- The harmonic average of Precision and Recall

$$Precision = \frac{tp}{tp + fp} = \frac{tp}{\text{Predicted as positive}}$$

$$Recall = \frac{tp}{tp + fn} = \frac{tp}{\text{Actually positive}}$$

$$F - score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Sensitivity (敏感度) and Specificity (特異度)

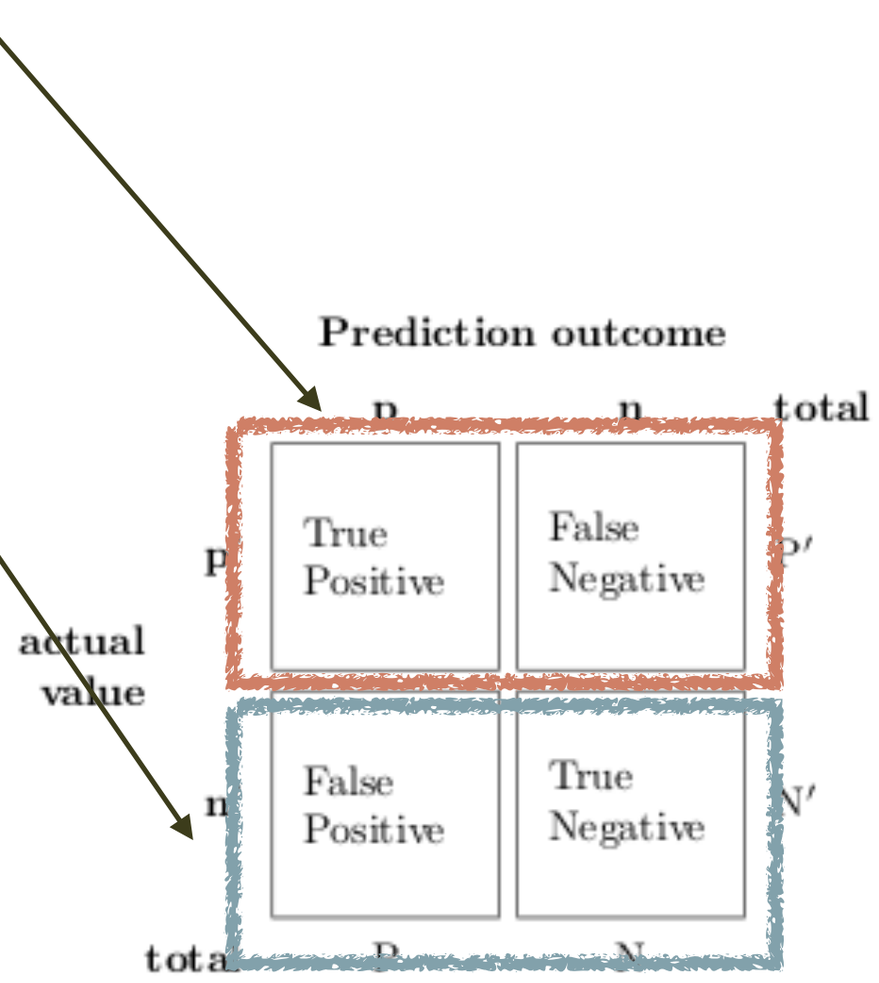
- Two metrics wide-used in the data mining and medical domains
- Sensitivity = Recall
  - How many positive cases can be confirmed
- Specificity = Recall of negative cases
  - How many negative cases can be confirmed

Sensitivity (敏感度) and Specificity (特異度)

$$Precision = \frac{tp}{tp + fp} = \frac{tp}{\text{Predicted as positive}}$$

$$Sensitivity = \frac{tp}{tp + fn} = \frac{tp}{\text{Actually positive}}$$

$$Specificity = \frac{tn}{tn + fp} = \frac{tn}{\text{Actually negative}}$$



Performance Evaluation for Multiway Classification

- A news classifier that classifies a news article into one of four categories.

- Compute the Precision, Recall, and F-Score in binary

Category	Precision	Recall	F-Score	
Education	38%	57%	46%	← Positive: Education Negative: Others
Finance	93%	83%	88%	← Positive: Finance Negative: Others
World	77%	87%	82%	← Positive: World Negative: Others
Society	87%	89%	88%	← Positive: Society Negative: Others

Confusion Matrix

- For multiway classification, confusion matrix provides useful information for understanding the behaviors of the model.

		Predicted Categories			
		Education	Finance	World	Society
Actual Categories	Education	**89**	15	8	20
	Finance	11	**96**	26	7
	World	15	23	**75**	3
	Society	45	14	24	**58**

Macro-Averaging

- Simply averaging

$$Macro\ F-Score = \frac{F_{education} + F_{finance} + F_{world} + F_{society}}{4}$$

$$= \frac{46\% + 88\% + 82\% + 88\%}{4}$$

Category	Precision	Recall	F-Score
Education	38%	57%	46%
Finiance	93%	83%	88%
World	77%	87%	82%
Society	87%	89%	88%
Marco-Averaged	74%	79%	76%

Micro-Averaging

- Weighted average according to the category size

$$Micro\ F-Score = \frac{F_{education} \times C_{education} + F_{finance} \times C_{finance} + F_{world} \times C_{world} + F_{society} \times C_{society}}{C_{education} + C_{finance} + C_{world} + C_{society}}$$

$$= \frac{46\% \times 10 + 88\% \times 60 + 82\% \times 450 + 88\% \times 270}{10 + 60 + 450 + 270}$$

Category	Precision	Recall	F-Score	# Instances
Education	38%	57%	46%	10
Finiance	93%	83%	88%	60
World	77%	87%	82%	450
Society	87%	89%	88%	270
Marco-Averaged	74%	79%	76%	
Micro-Averaged	**81%**	**87%**	**84%**	

Macro Averaging vs Micro Averaging

- Macro-averaging
  - Giving equal weight to each category in spite of its occurrences
- Micro-averaging
  - Giving equal weight to each instance
  - Dominated by the large categories
  - Micro F-score ~ Accuracy

Category	Precision	Recall	F-Score	# Instances
Education	38%	57%	46%	10
Finiance	93%	83%	88%	60
World	77%	87%	82%	450
Society	87%	89%	88%	270
Marco-Averaged	74%	79%	76%	
Micro-Averaged	81%	87%	84%	

How to Confirm the Performance Difference

- Is the model B really superior to the model A?
  - or just accidentally?

Model	Accuracy	Precision	Recall	F1
A	82.91%	88.51%	76.24%	81.91%
B	**83.42%**	**89.66%**	**76.47%**	**82.54%**

How to Confirm the Performance Difference

- B does only one more correct prediction over A

Model	Accuracy	Precision	Recall	F1
A	82.91%	88.51%	76.24%	81.91%
B	**83.42%**	**89.66%**	**76.47%**	**82.54%**

A	Predicted as Positive	Predicted as Negative
Actually Positive	77	24
Actually Negative	10	88

B	Predicted as Positive	Predicted as Negative
Actually Positive	**78**	24
Actually Negative	**9**	88

How to Confirm the Performance Difference

- Does C significantly outperform A and B?

Model	Accuracy	Precision	Recall	F1
A	82.91%	88.51%	76.24%	81.91%
B	83.42%	89.66%	76.47%	82.54%
C	**89.16%**	**92.08%**	**86.92%**	**89.42%**

A	Predicted Positive	Predicted Negative
Actually Positive	77	24
Actually Negative	10	88

B	Predicted Positive	Predicted Negative
Actually Positive	78	24
Actually Negative	9	88

B	Predicted Positive	Predicted Negative
Actually Positive	**93**	**14**
Actually Negative	**8**	88

Substance of a Performance Difference

- The difference that is making sense but not by chance depends on
  - The difference gap
    - 1% is subtle
    - 10% is more substantial
  - The number of test cases
    - Different in 5 samples is more likely accident
    - Different in 1,000 samples is much more representative
- Do we have a scientific way to measure the substance of a performance difference?

Statistical Significance Tests

- Designed to quantify how unlikely the result to have occurred given the null hypothesis.
    - If the significance test **rejects** the null hypothesis, the result is statistically significant.
- Can improve the confidence in the selection and the interpretation of models.
- Can aid in comparing models and choosing a final one.

https://machinelearningmastery.com/statistical-significance-tests-for-comparing-machine-learning-algorithms/

Statistical Hypothesis Tests

- Comparing samples quantifies how likely it is to observe two data samples given the assumption that the samples have the same distribution.
    - Null hypothesis
- If the test suggests that evidence is not enough to reject the null hypothesis
    - The performance difference is likely due to statistical **chance**.
- If the test suggests that evidence is enough to reject the null hypothesis
    - The performance difference is likely due to a real difference in the models.

McNemar's Test

- The null hypothesis is the two models have the same error rate
  - $n_{01} = n_{10}$
- Chi-square for checking if the null hypothesis is rejected

$$\frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$$

	Misclassified by A	Corrected classified by A
Misclassified by B	n_{00}	n_{01}
Corrected classified by B	n_{10}	n_{11}

vs

	Misclassified by A	Corrected classified by A
Misclassified by B	n_{00}	$(n_{01}+n_{10}) / 2$
Corrected classified by B	$(n_{01}+n_{10}) / 2$	n_{11}

McNemar's Test

- Accuracy of A: (85+9)/(20+9+6+85) = 0.783333333
- Accuracy of B: (85+6)/(20+9+6+85) = 0.758333333
- Does A outperform B significantly?

	Misclassified by A	Corrected classified by A
Misclassified by B	20	9
Corrected classified by B	6	85

McNemar's Test

- Accuracy of A: (85+9)/(20+9+6+85) = 0.783333333
- Accuracy of B: (85+6)/(20+9+6+85) = 0.758333333
- A does not significantly outperform B at p=0.05

$$\frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} = \frac{(|9 - 6| - 1)^2}{6 + 9}$$
$$= \frac{4}{15} = 0.266666667 < 3.841459$$

	Misclassified by A	Corrected classified by A
Misclassified by B	20	9
Corrected classified by B	6	85

McNemar's Test

- Accuracy of A: (85+9)/(20+9+6+85) = 0.783333333
- Accuracy of B: (85+6)/(20+9+6+85) = 0.758333333
- A significantly outperforms B at p=0.05

$$\frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} = \frac{(|13 - 2| - 1)^2}{13 + 2}$$
$$= \frac{121}{14} = 8.64285714 > 3.841459$$

	Misclassified by A	Corrected classified by A
Misclassified by B	20	13
Corrected classified by B	2	85

Word Sense Disambiguation

Word Senses

- Many words have several meanings or senses
- Bank
  - The rising ground bordering a lake, river, or sea (河岸、湖岸、海岸)
  - An establishment for the custody, loan exchange, or issue of money, for the extension of credit, and for facilitating the transmission of funds （銀行）

Even More Senses Defined …

WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: bank Search WordNet

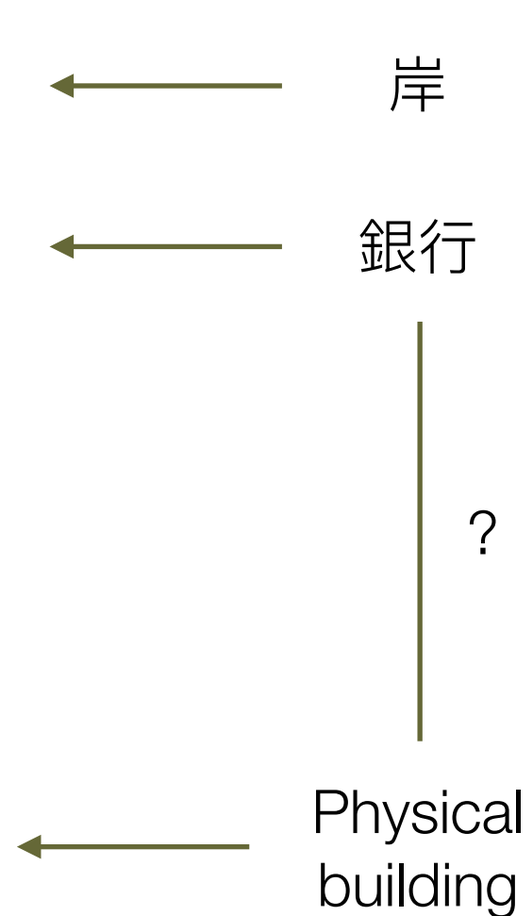
Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

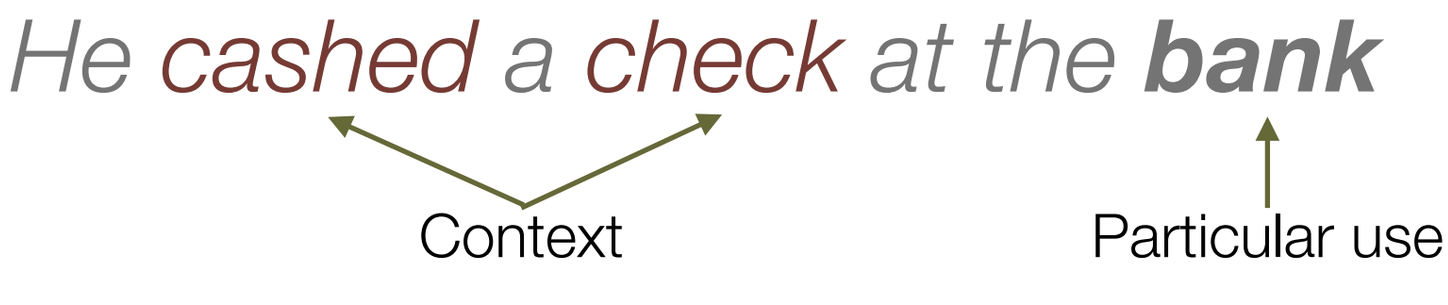
- S: (n) **bank** (sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
- S: (n) depository financial institution, **bank**, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending activities) *"he cashed a check at the bank"; "that bank holds the mortgage on my home"*
- S: (n) **bank** (a long ridge or pile) *"a huge bank of earth"*
- S: (n) **bank** (an arrangement of similar objects in a row or in tiers) *"he operated a bank of switches"*
- S: (n) **bank** (a supply or stock held in reserve for future use (especially in emergencies))
- S: (n) **bank** (the funds held by a gambling house or the dealer in some gambling games) *"he tried to break the bank at Monte Carlo"*
- S: (n) **bank**, cant, camber (a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force)
- S: (n) savings bank, coin bank, money box, **bank** (a container (usually with a slot in the top) for keeping money at home) *"the coin bank was empty"*
- S: (n) **bank**, bank building (a building in which the business of banking transacted) *"the bank is on the corner of Nassau and Witherspoon"*



Word Sense Disambiguation

- To determine which of the senses of an ambiguous word is invoked in a particular use of the word.

- By looking the context of the word.



Word Senses and Word Sense Disambiguation

- A word is assumed to have a finite number of senses
  - Defined by a dictionary or other resources
- Word sense disambiguation
  - Predict the most suitable sense from a number of senses predefined.

Most Ambiguous Words

Word	Number of Senses
break	75
broken	72
cut	70
broke	60
breaking	60
run	57
cutting	54
made	52
running	52
making	52
play	52
make	51
better	50
giving	48

Word Sense Disambiguation

- Given
  - A target word *w* in context
  - A infinite set of the senses, S_w, of *w*
- Determine the sense s' of *w* from S_w

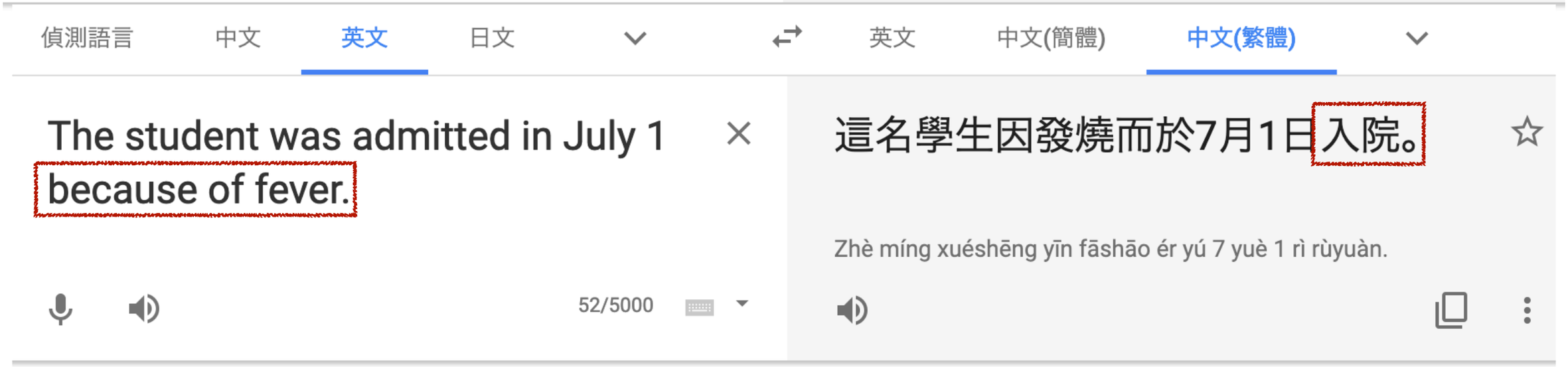
$$s' = \arg\max_{s \in S_w} P(s|w, C)$$

where w is the target word,
C is the contextual information of w,
and S_w is the senses of w

Applications

- Machine translation
- Question answering
- Speech Synthesis
- 種 (種 花 / 品 種)

WSD for Machine Translation



Approaches to Word Sense Disambiguation

- Supervised
  - Requiring a corpus on which word senses are labelled.
  - The supervised classifier can be trained to perform WSD.
- Unsupervised
  - Only a lexicon with senses for each word is available.
  - No labeled data can be used for training the model.
- Semi-supervised

Supervised Machine Learning

- A labeled corpus
  - A corpus consists of words that are labeled with their sense.
- And we can train a model to capture the relationship between sense, word, and the context

$$P(s|w, C)$$

- Finally, we can predict the sense of a word by using

$$s' = \arg\max_{s \in S_w} P(s|w, C)$$

Bayes' Theorem

- Given an instance represented by a vector ***x*** representing *m* features, it assigns to this instance probabilities for each of categories.

$$P(y|\mathbf{x}) = P(y|x_1, x_2, x_3, ..., x_m)$$

a little more difficult to calculate

- Bayes' Theorem

Frequency of a news category; easier to calculate

$$P(y|\mathbf{x}) = \frac{P(y)P(\mathbf{x}|y)}{P(\mathbf{x})}$$

still difficult to calculate

easier to calculate

The “Naive” Assumption

- We have a strong (naive) assumption that all the features x_1, x_2, x_3, …, x_n are independent.

- So the calculation can be easier

difficult to calculate

$$P(\mathbf{x}|y) = P(x_1, x_2, x_3, ..., x_n|y)$$
$$= P(x_1|y)P(x_2|y)P(x_3|y)...P(x_n|y)$$
$$= \prod_{i=1}^{n} P(x_i|y)$$

easier to calculate

Stands only if all x_i are independent to each other

Naive Bayes Classifier

- So we can predict the label *y'* given the features in ***x***

$$P(y|\mathbf{x}) = P(y)\prod_{i=1}^{m}\frac{P(x_i|y)}{P(x_i)} \propto \log P(y) + \sum_{i=1}^{m}\log P(x_i|y)$$

$$y' = \arg\max_{y\in C} P(y)\prod_{i=1}^{m}\frac{P(x_i|y)}{P(x_i)}$$

$$= \arg\max_{y\in C} \log P(y) + \sum_{i=1}^{m}\log P(x_i|y)$$

Add-One Smoothing for OOV

- $P(x_i|y) = 0$ if x_i is unseen.
- Like add-one smoothing used in language modeling

$$P(x_i|y) = \frac{C(x_i, y) + 1}{C(y) + |V|}$$

Statistically Independence

- Event ***a*** and event ***b*** are independent if and only if their joint probability equals to the product of their probabilities.

$$P(A, B) = P(A)P(B)$$

- No correlation between events a and b

Counterexample of Statistically Independent

- Event a: woman
  - P(a) = 0.5
- Event b: breast cancer (乳癌)
  - P(b) = 0.001
- The probability of a woman with breast cancer
  - P(a,b) = P(a) * P(b) = 0.0005

Naive Bayes Classifier vs Language Model Classifier

Probability of a good or a bad review;
We ignored this since the ratio is 1:1 in the corpus

Probability of a review according to a language model of either bad reviews or good reviews

$$P(y|\mathbf{x}) = \frac{P(y)P(\mathbf{x}|y)}{P(\mathbf{x})}$$

We ignored this because x is the same for a particular input

Supervised Model for WSD

- Training the Naive Bayes model

foreach *sense* s_k *of* w **do**
  $\mathrm{P}(\mathrm{s}_k) \leftarrow \frac{C(s_k)}{C(w)}$
  foreach *word* v *in the vocabulary* **do**
    $\mathrm{P}(\mathrm{v}|\mathrm{s}_k) \leftarrow \frac{C(v,s_k)}{C(v)}$
  end
end

Supervised Model for WSD

- Predicting the sense of w given its contextual words

foreach $sense\ s_k\ of\ w$ **do**
  score(s_k) $\leftarrow logP(s_k)$
  foreach *word* v *in the context of* w **do**
    score(s_k) $\leftarrow score(s_k) + logP(v|s_k)$
  end
end

Unsupervised Approach

- Lesk Algorithm (1986)
- Simplified Lesk Algorithm (2004)
- Basic idea is to compute the similarity of the context of *w* and the definition of *w*.

Definition of the first sense of bank

bank[1] /bæŋk/ ●●● S1 W1 **noun** [countable]

1 PLACE FOR MONEY

a) a business that keeps and lends money and provides other financial services

in the bank

- We have very little money in the bank.
- Barclays Bank
- a **bank loan**

b) a local office of a bank

- I have to **go to the bank** at lunch time.

→ **clearing bank, merchant bank**

He deposited the money at the ***bank***

Idea of Lesk Algorithm

- Counting the amount of words that are in both context of that target word *w* and in the definition of that sense in a dictionary.

- The outcome is the sense which has the largest number of this count.

Lesk Algorithm

$$\textbf{foreach } sense\ s_k\ of\ w \textbf{ do}$$
$$\quad \text{score}(s_k) \leftarrow overlap(D_k, \cup_{v \in c} E_v)$$
$$\textbf{end}$$

Words used to define s_k

All words used to define v

- This cigar **burns** slowly and creates a stiff ***ash***
- The ***ash*** is one of the last **trees** to come into leaf

s_1	ash	tree (白蠟木)	a tree of the olive family
s_2	ash	burned stuff (灰燼)	the solid residue left when combustible material is burned

Lesk Algorithm (Simplified)

```
def wsd(word, context):
    outcome = None
    largest = 0
    for s in wordnet.synsets(word):
        definition_words = s.definition().split()
        overlap = len(context & definition_words)
        if overlap > largest:
            largest = overlap
            outcome = s
return outcome
```

Lesk Algorithm

foreach *sense* s_k *of* w **do**
$\quad \text{score}(s_k) \leftarrow overlap(D_k, \cup_{v \in c} E_v)$
end

```
def wsd(word, context):
    outcome = None
    largest = 0

    exp_context = set()
    for v in context:
        for s in wordnet.synsets(v):
            exp_context |= set(s.definition().split())

for s in wordnet.synsets(word):
    definition_words = s.definition().split()
    overlap = len(exp_context & definition_words)
    if overlap > largest:
        largest = overlap
        outcome = s
return outcome
```

Example 1

- bank
  - **bank.n.01**: “sloping land (especially the slope beside a body of water)”
  - **depository_financial_institution.n.01**: “a financial institution that accepts deposits and channels the money into lending activities”
  - **savings_bank.n.02**: a container (usually with a slot in the top) for keeping money at home
  -

*He sat on the **bank** of the river and watched the currents*

Example 2

- made (up to 52 seneses)
  - engage in
  - give certain properties to something
  - make or cause to be or to become
  - cause to do; cause to act in a specified manner
  - give rise to; cause to happen or occur, not always intentionally
  - create or manufacture a man-made product
  - make, formulate, or derive in the mind
  - compel or make somebody or something to act in a certain way
  - create by artistic means
  - earn on some commercial or business transaction; earn as salary or wages
  - ....
  - **Form by assembling individuals or constituents**

*This chair was **made** by my friend*

Thesaurus-based Disambiguation

- Exploiting the semantic categorization provided by a thesaurus (同義辭辭典) like Roget's.

- The semantic categories of the words in a context determine the semantic category of the context as a whole.

- This category can determine which word senses are used.

Thesaurus

adj. **at the mercy of; answerable** | *noun* **issue, matter** | *noun* **one under authority of another**

Paste a sentence with the word “bank” to see some examples in-context.

SHOW EXAMPLES

subject

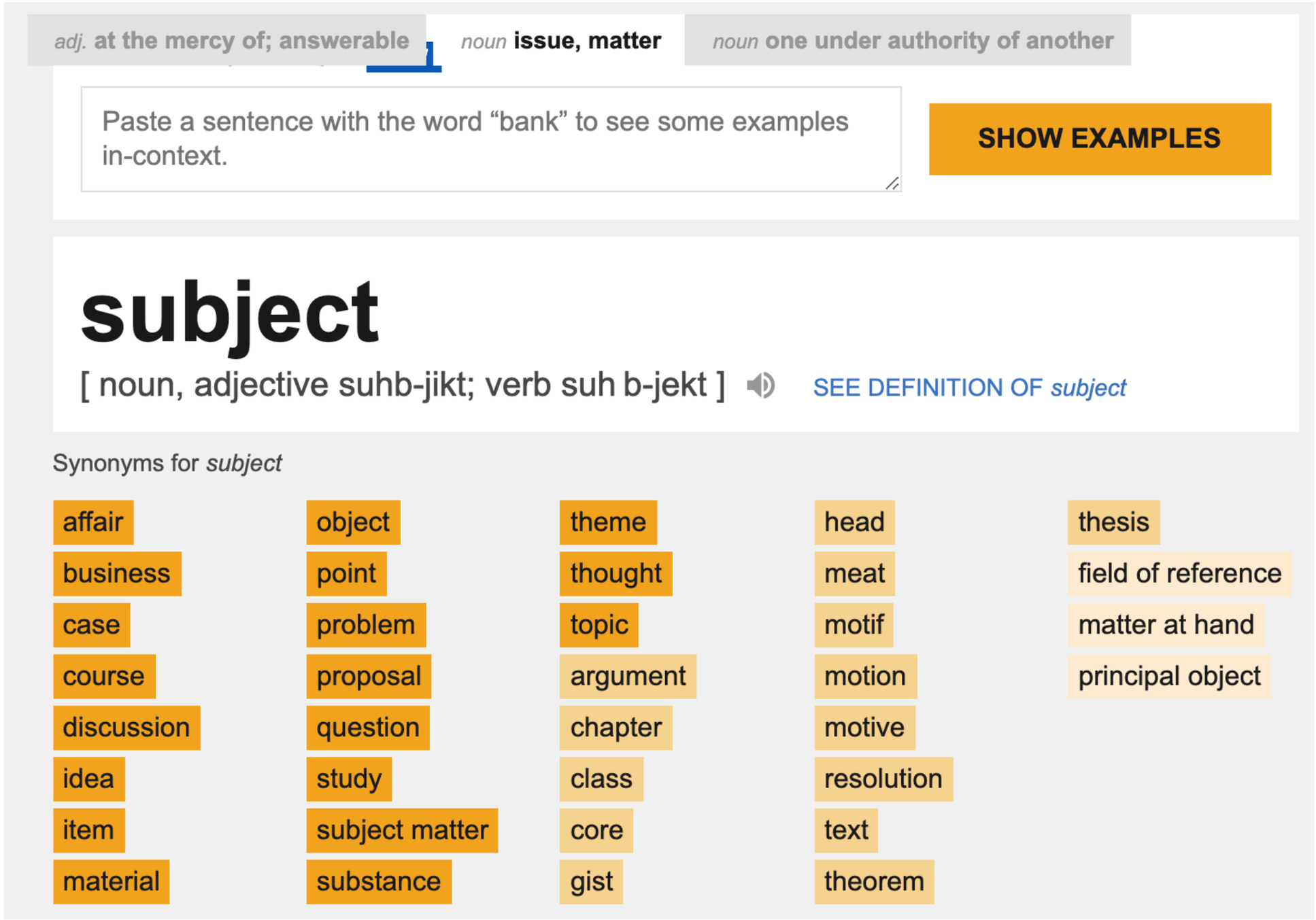
[noun, adjective suhb-jikt; verb suh b-jekt] SEE DEFINITION OF *subject*

Synonyms for *subject*

accountable	secondary	disposed	subordinate	submissive
apt	sensitive	enslaved	tributary	subservient
conditional	susceptible	governed	at one's feet	substract
dependent	vulnerable	open	bound by	tentative
exposed	captive	ruled	in danger of	under
inferior	collateral	satellite	obedient	
liable	contingent	sub	provisional	
likely	controlled	subaltern	servile	
prone	directed	subjugated	slavish	

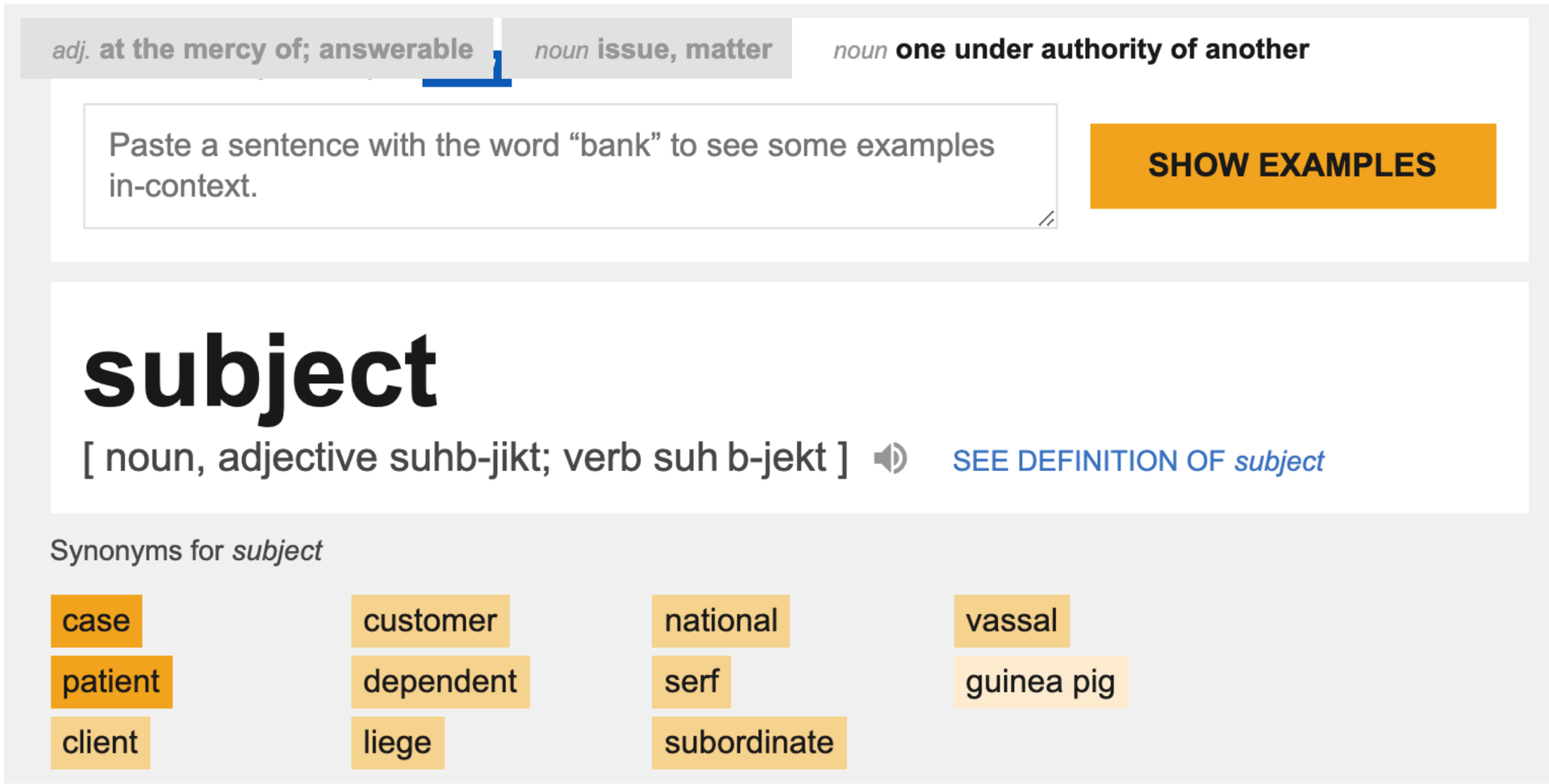
https://www.thesaurus.com/browse/subject?s=t

Thesaurus



https://www.thesaurus.com/browse/subject?s=t

Thesaurus



https://www.thesaurus.com/browse/subject?s=t

Idea of Walker's (1987) Algorithm

- If the word is assigned several categories, then we assume these categories correspond to the different senses of the word.

  - $t(s_k)$ is the category of sense s_k of ambiguous word *w* given context *c.*

  - *w* can be disambiguated by counting the number of words for which the thesaurus lists $t(s_k)$ *as possible category.*

Thesaurus-based WSD

foreach *sense* s_k *of* w **do**
    $\text{score}(\text{s}_k) \leftarrow 0$
    foreach *word* v *in the context of* w **do**
        if $t(s_k)$ *is one of the category of* v **then**
            $\text{score}(\text{s}_k) \leftarrow score(s_k) + 1;$
        end
    end
end

Drawback of Lesk Algorithm

- Very sensitive to the exact wording of definitions
    - The performance may be greatly reduced when some important words missing from the context.
- Sensitive to how the dictionary is composed.
    - Dictionary glosses are usually short and do not provide sufficient vocabulary to relate fine-grained sense distinctions.

Supervised and Unsupervised

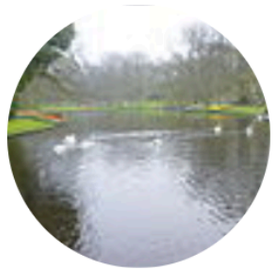
	Pros	Cons
Superivsed	• Higher performance	• Requiring labeled data • Lower coverage • Sensitive to the training data
Unsupervised	• High coverage (limited by the dictionary)	• Generally poorer performance • Sensitive to the dictionary

Babelfy

- Multilingual Word Sense Disambiguation and Entity Linking

He **sat** on the **bank** of the **river** and **watched** the **currents**

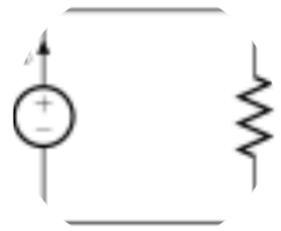
EN baby-sit
EN Work or act as a baby-sitter



河 岸
EN Sloping land (especially the slope beside a body of water)

河
EN A large natural stream of water (larger than a creek)

EN watch
EN Follow with the eyes or the mind



电 流
EN A flow of electricity through a conductor

Multilingual Supportive of Babelfy

He cashed a check at the bank

EN cash in
EN Exchange for cash

支票
EN A written order directing a bank to pay money

銀行
金融机构

我 从 政治 大学 取得 硕士 学位

國立政治大學
EN National Chengchi University is a public university located in Taipei, Taiwan.

硕 士 学 位
EN An academic degree higher than a bachelor's degree but lower than a doctor's degree

ceived funding from the Euro under grant agreer
pt and design by Andrea Mo
lfy and its API are li

Labeled Word Corpus

- https://dkpro.github.io/dkpro-wsd/corpora/

Table of WSD corpora

Date	Corpus	Language	Style	Train	Format	Inventory	POS	Lemma	Sent.	Notes
1998	Senseval-1 task: English lexical sample	en	LS	no	Senseval2LS	HECTOR	?	?	?	1
2001	Senseval-2 task: Basque lexical sample	eu	LS	no	Senseval2LS	Euskal Hiztegia subset, TEI-SGML	yes	yes	no	
2001	Senseval-2 task: Czech all words	cs	AW	no	Senseval2AW	custom, text	no	no	no	
2001	Senseval-2 task: Dutch all words	nl	AW	no	Senseval2AW	none	no	no	yes	
2001	Senseval-2 task: English all words	en	AW	no	Senseval2AW	WordNet 1.7pre	no	no	no	2
2001	Senseval-2 task: English all words (Rada Mihalcea's conversion)	en	AW	no	SemCor	WordNet 1.7.1 through 3.0	yes	yes	yes	3
2001	Senseval-2 task: English group lexical sample	en	LS	yes	Senseval2LS	WordNet 1.7pre	yes	yes	no	2
2001	Senseval-2 task: English lexical sample	en	LS	no	Senseval2LS	WordNet 1.7pre	yes	yes	no	2, 3
2001	Senseval-2 task: Estonian all words	et	AW	yes	Senseval2AW	Estonian EWN v37	no	no	no	

SemEval: Semantic Evaluation

- A shared-task/workshop for natural language processing since 1998
- Exploring the latest progress of NLP
- Interesting tasks in SemEval 2019
  - Hyperpartisan (極左、極右) News Detection
  - Detection of Hate Speech
  - Identifying offensive language
  - Rumor detection

SemEval-2019

International Workshop on Semantic Evaluation

Sponsored by SIGLEX and Microsoft

Tasks

We are pleased to announce the following tasks in SemEval-2019.

Frame semantics and semantic parsing

- Task 1: Cross-lingual Semantic Parsing with UCCA [mailing list] [email organizers]
- Task 2: Unsupervised Lexical Semantic Frame Induction [mailing list] [email organizers]

Opinion, emotion and abusive language detection

- Task 3: EmoContext: Contextual Emotion Detection in Text [discussion group] [email organizers]
- Task 4: Hyperpartisan News Detection [mailing list] [email organizers]
- Task 5: HatEval: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter [mailing list] [email organizers]
- Task 6: OffensEval: Identifying and Categorizing Offensive Language in Social Media [mailing list] [email organizers]

Fact vs fiction

- Task 7: RumourEval 2019: Determining Rumour Veracity and Support for Rumours [discussion group] [email organizers]
- Task 8: Fact Checking in Community Question Answering Forums [mailing list] [email organizers]

Information extraction and question answering

- Task 9: Suggestion Mining from Online Reviews and Forums [mailing list] [email organizers]
- Task 10: Math Question Answering [mailing list] [email organizers]

Contact Info

Organizers

- Jonathan May, ISI, University of Southern California
- Ekaterina Shutova, University of Amsterdam
- Aurelie Herbelot, University of Trento
- Xiaodan Zhu, Queen's University
- Marianna Apidianaki, LIMSI, CNRS, Université Paris-Saclay & University of Pennsylvania
- Saif M. Mohammad, National Research Council Canada

Email

semeval-organizers@googlegroups.com

Note that this is the mailing list for SemEval organizers. For questions on a particular task, post them at the *task* mailing list or contact the task organizers directly. You can find the task mailing list from the task webpage.

Other Info

Announcements

- 2019/2/17: We are excited to announce that Sam Bowman from NYU will give a keynote at the workshop. Please click here for details

http://alt.qcri.org/semeval2019/index.php?id=tasks