Information Extraction

Lecture 4 – Named Entity Recognition II

CIS, LMU München
Winter Semester 2015-2016

Dr. Alexander Fraser, CIS

Administravia

- How many people need a Seminar topic who have not yet registered?
  - You must register by Thursday evening!
  - Please also check the web page to make sure I recorded your topic correctly!

- And now for something completely different:
  - How many people took Höhere Programmierung?
  - How many people took Morphology?

Reading

- Please read Sarawagi Chapter 3 for next time
    - Sarawagi talks about classifier based IE in Chapter 3
    - Unfortunately, the discussion is very technical. I would recommend reading it, but not worrying too much about the math (yet), just get the basic idea
    - You may find yourself wanting to reread Chapter 3 again after we discuss machine learning

Back to the Future

- We'll start by completing the slide set from last week (evaluation in IE)

Outline

- Evaluation in more detail
  - Look at Information Retrieval
- Return to Rule-Based NER
  - The CMU Seminar dataset
- Issues in Evaluation of IE
- Human Annotation for NER

Recall

Measure of how much relevant information the system has extracted (coverage of system).

Exact definition:

Recall = 1 if no possible correct answers

else:

$$\frac{\text{\# of correct answers given by system}}{\text{total \# of possible correct answers in text}}$$

Slide modified from Butt/Jurafsky/Martin

Precision

Measure of how much of the information the system returned is correct (accuracy).

Exact definition:

Precision = 1 if no answers given by system

else:

$$\frac{\text{\# of correct answers given by system}}{\text{\# of answers given by system}}$$

Slide modified from Butt/Jurafsky/Martin

Evaluation

Every system, algorithm or theory should be **evaluated**, i.e. its output should be compared to the **gold standard** (i.e. the ideal output). Suppose we try to find scientists…

Algorithm output:

O = {Einstein, Bohr, Planck, Clinton, Obama}

✓ ✓ ✓ ✗ ✗

Gold standard:

G = {Einstein, Bohr, Planck, Heisenberg}

✓ ✓ ✓ ✗

Precision:

What proportion of the output is correct?

$$\frac{|\, O \wedge G \,|}{|O|}$$

Recall:

What proportion of the gold standard did we get?

$$\frac{|\, O \wedge G \,|}{|G|}$$

Slide modified from Suchanek

Evaluation

- Why Evaluate?
- What to Evaluate?
- How to Evaluate?

Slide from Giles

Why Evaluate?

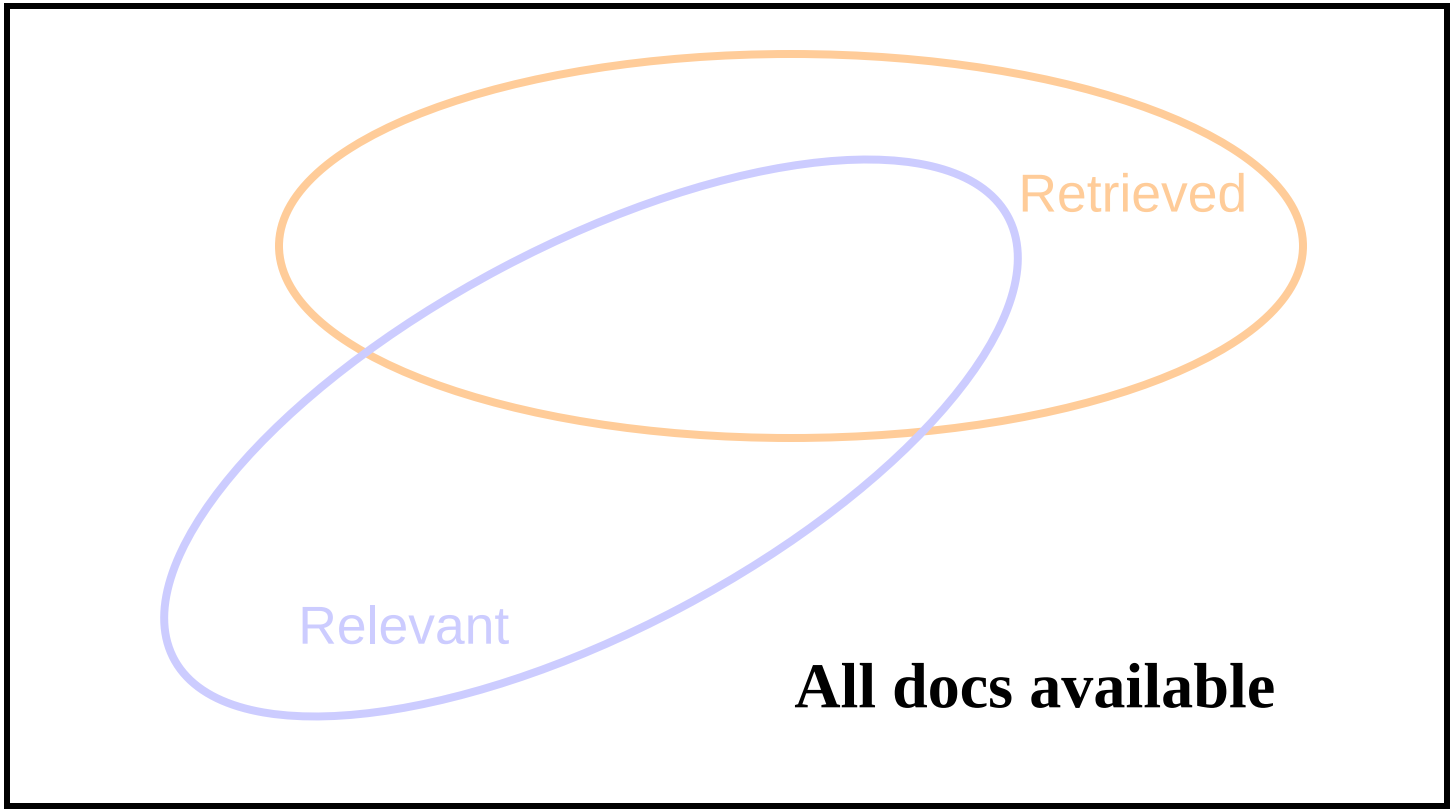
- Determine if the system is useful
- Make comparative assessments with other methods/systems
  - Who’s the best?
- **Test and improve systems**
- Others: Marketing, …

Slide modified from Giles

What to Evaluate?

- In Information Extraction, we try to match a pre-annotated gold standard
- But the evaluation methodology is mostly taken from Information Retrieval
  - So let's consider **relevant documents** to a search engine **query** for now
  - We will return to IE evaluation later

Relevant vs. Retrieved Documents



Set approach

Slide from Giles

Contingency table of relevant and retrieved documents

		relevant	
		Rel	NotRel
retrieved	Ret	Ret_{Rel}	$\text{Ret}_{\text{NotRel}}$
	NotRet	$\text{NotRet}_{\text{Rel}}$	$\text{NotRet}_{\text{NotRel}}$

$\text{Ret} = \text{Ret}_{\text{Rel}} + \text{Ret}_{\text{NotRel}}$

$\text{NotRet} = \text{NotRet}_{\text{Rel}} + \text{NotRet}_{\text{NotRel}}$

$\text{Relevant} = \text{Ret}_{\text{Rel}} + \text{NotRet}_{\text{Rel}}$

$\text{Not Relevant} = \text{Ret}_{\text{NotRel}} + \text{NotRet}_{\text{NotRel}}$

Total # of documents available $N = \text{Ret}_{\text{Rel}} + \text{NotRet}_{\text{Rel}} + \text{Ret}_{\text{NotRel}} + \text{NotRet}_{\text{NotRel}}$

- Precision: $P = \text{Ret}_{\text{Rel}} / \text{Retrieved}$
- Recall: $R = \text{Ret}_{\text{Rel}} / \text{Relevant}$

$P = [0,1]$

$R = [0,1]$

Slide from Giles

Contingency table of classification of documents

		Actual Condition	
		Present	Absent
Test result	Positive	tp	fp type1
	Negative	fn type2	tn

fp type 1 error

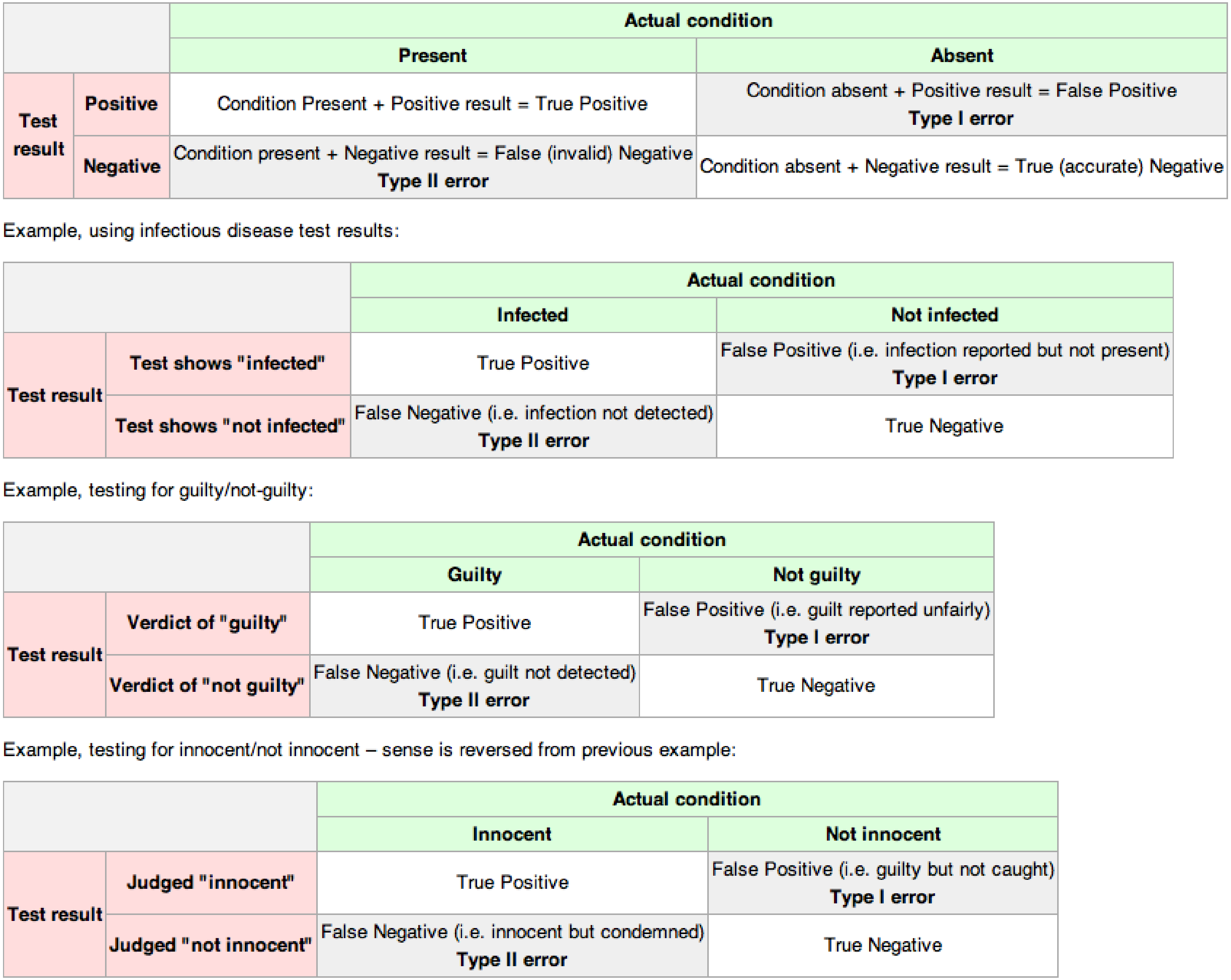
fn type 2 error

present = tp + fn
positives = tp + fp
negatives = fn + tn

Total # of cases N = tp + fp + fn + tn

- False positive rate α = fp/(negatives)
- False negative rate β = fn/(positives)

Slide from Giles



		Actual condition	
		Present	**Absent**
Test result	**Positive**	Condition Present + Positive result = True Positive	Condition absent + Positive result = False Positive **Type I error**
	Negative	Condition present + Negative result = False (invalid) Negative **Type II error**	Condition absent + Negative result = True (accurate) Negative

Example, using infectious disease test results:

		Actual condition	
		Infected	**Not infected**
Test result	**Test shows "infected"**	True Positive	False Positive (i.e. infection reported but not present) **Type I error**
	Test shows "not infected"	False Negative (i.e. infection not detected) **Type II error**	True Negative

Example, testing for guilty/not-guilty:

		Actual condition	
		Guilty	**Not guilty**
Test result	**Verdict of "guilty"**	True Positive	False Positive (i.e. guilt reported unfairly) **Type I error**
	Verdict of "not guilty"	False Negative (i.e. guilt not detected) **Type II error**	True Negative

Example, testing for innocent/not innocent – sense is reversed from previous example:

		Actual condition	
		Innocent	**Not innocent**
Test result	**Judged "innocent"**	True Positive	False Positive (i.e. guilty but not caught) **Type I error**
	Judged "not innocent"	False Negative (i.e. innocent but condemned) **Type II error**	True Negative

Slide from Giles

Retrieval example

- Documents available: D1,D2,D3,D4,D5,D6, D7,D8,D9,D10
- Relevant: D1, D4, D5, D8, D10
- Query to search engine retrieves: D2, D4, D5, D6, D8, D9

	relevant	not relevant
retrieved		
not retrieved		

Slide from Giles

Retrieval example

- Documents available: D1,D2,D3,D4,D5,D6, D7,D8,D9,D10
- Relevant: D1, D4, D5, D8, D10
- Query to search engine retrieves: D2, D4, D5, D6, D8, D9

	relevant	not relevant
retrieved	D4,D5,D8	D2,D6,D9
not retrieved	D1,D10	D3,D7

Slide from Giles

Contingency table of relevant and retrieved documents

		relevant	
		Rel	NotRel
retrieved	Ret	$\text{Ret}_{\text{Rel}}=3$	$\text{Ret}_{\text{NotRel}}=3$
	NotRet	$\text{NotRet}_{\text{Rel}}=2$	$\text{NotRet}_{\text{NotRel}}=2$

$\text{Ret} = \text{Ret}_{\text{Rel}} + \text{Ret}_{\text{NotRel}}$
$= 3 + 3 = 6$

$\text{NotRet} = \text{NotRet}_{\text{Rel}} + \text{NotRet}_{\text{NotRe}}$
$= 2 + 2 = 4$

$\text{Relevant} = \text{Ret}_{\text{Rel}} + \text{NotRet}_{\text{Rel}}$
$= 3 + 2 = 5$

$\text{Not Relevant} = \text{Ret}_{\text{NotRel}} + \text{NotRet}_{\text{NotRel}}$
$= 2 + 2 = 4$

Total # of docs $N = \text{Ret}_{\text{Rel}} + \text{NotRet}_{\text{Rel}} + \text{Ret}_{\text{NotRel}} + \text{NotRet}_{\text{NotRel}} = 10$

- Precision: $P = \text{Ret}_{\text{Rel}}$ / Retrieved = 3/6 = .5
- Recall: $R = \text{Ret}_{\text{Rel}}$ / Relevant = 3/5 = .6

$P = [0,1]$
$R = [0,1]$

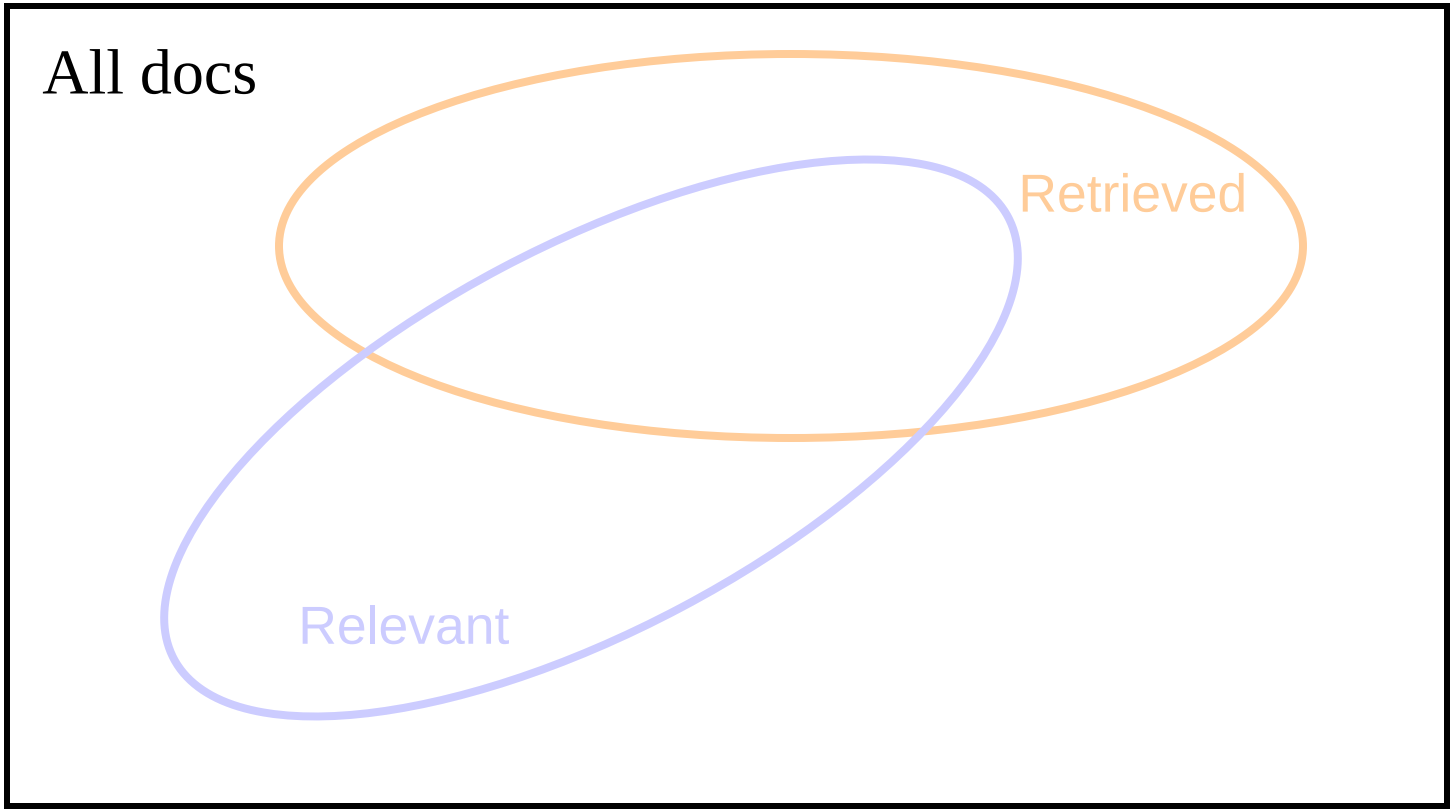
Slide from Giles

What do we want

- Find everything relevant – high recall
- Only retrieve what is relevant – high precision

Slide from Giles

Relevant vs. Retrieved

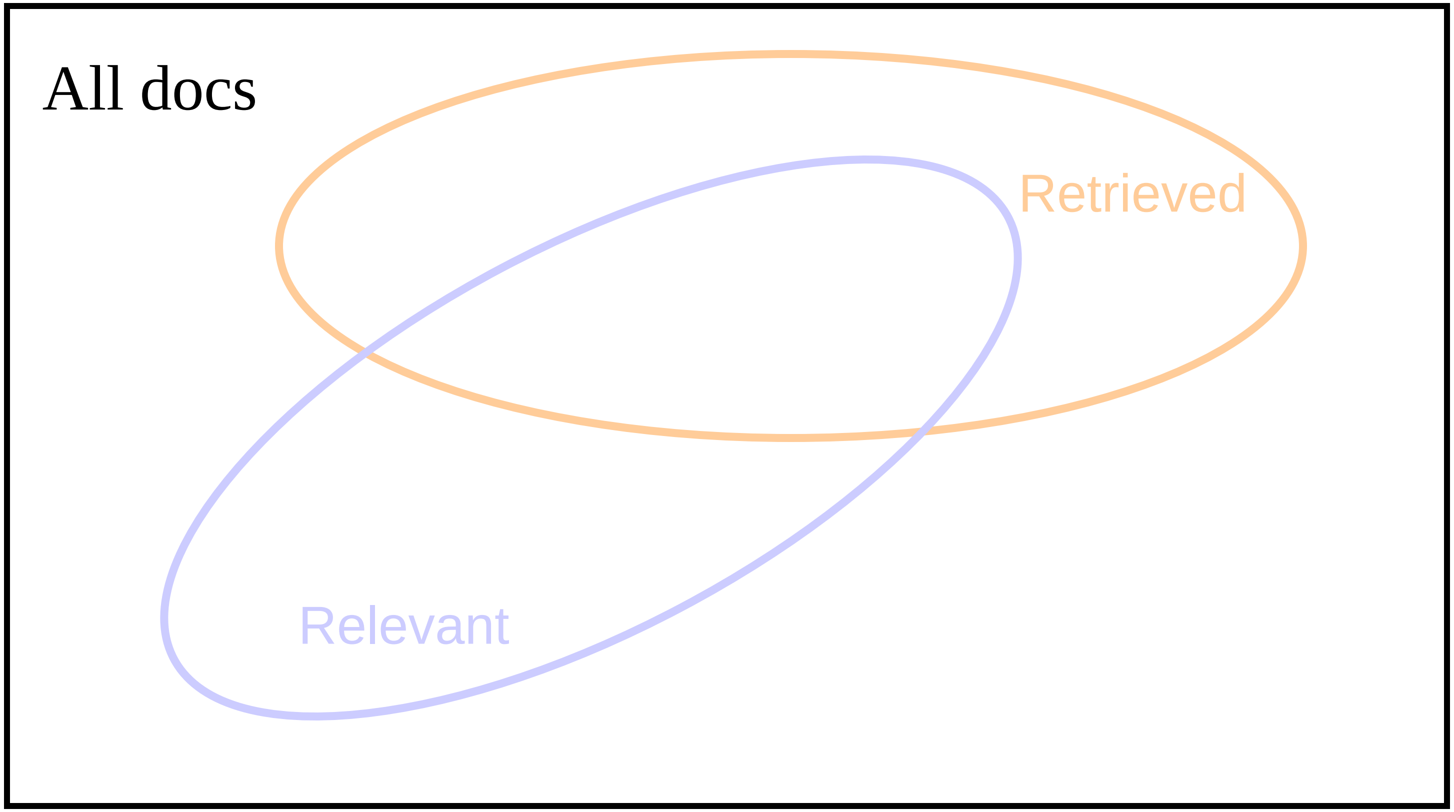


Slide from Giles

Precision vs. Recall

$$\text{Precision} = \frac{|\text{RelRetrieved}|}{|\text{Retrieved}|}$$

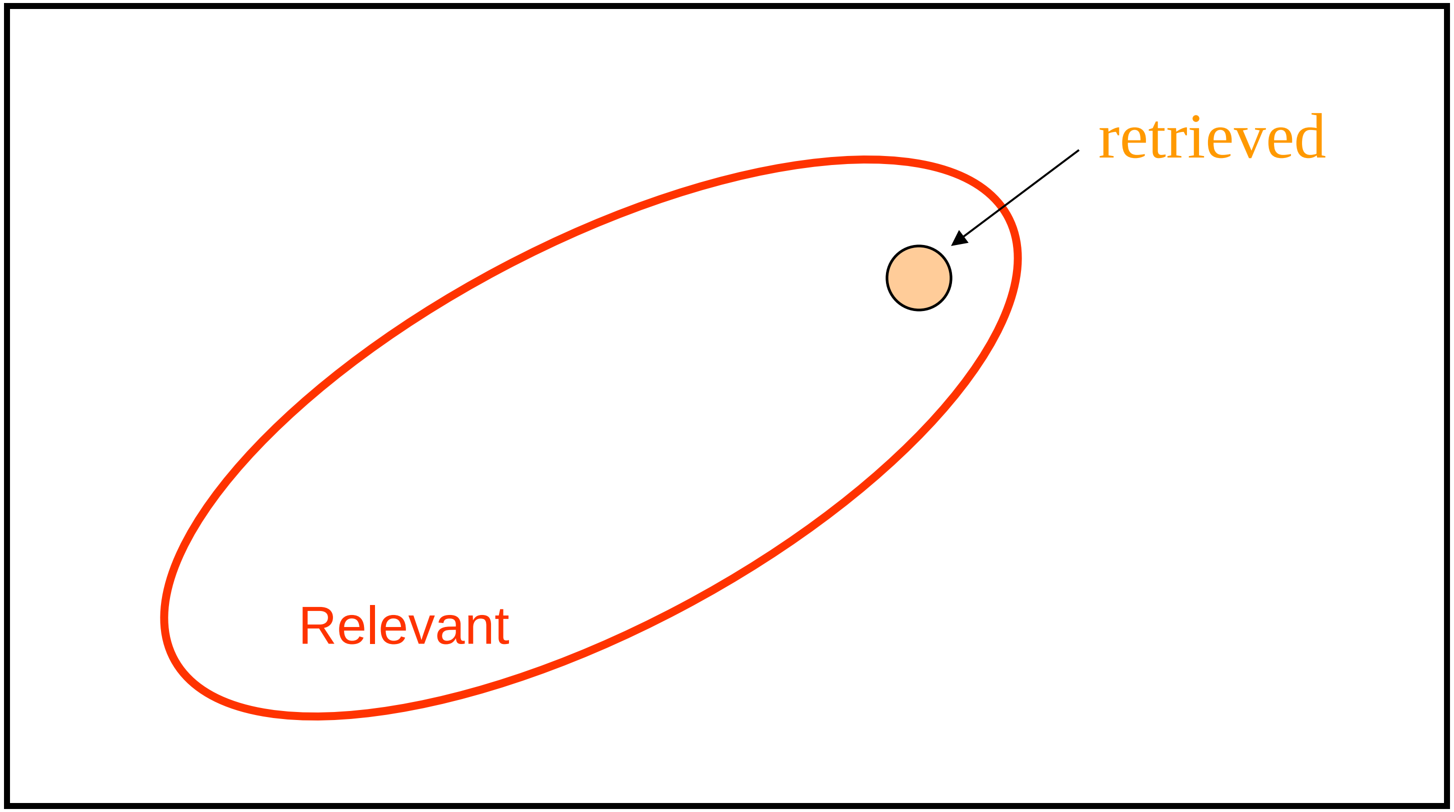
$$\text{Recall} = \frac{|\text{RelRetrieved}|}{|\text{Rel in Collection}|}$$



Slide from Giles

Retrieved vs. Relevant Documents

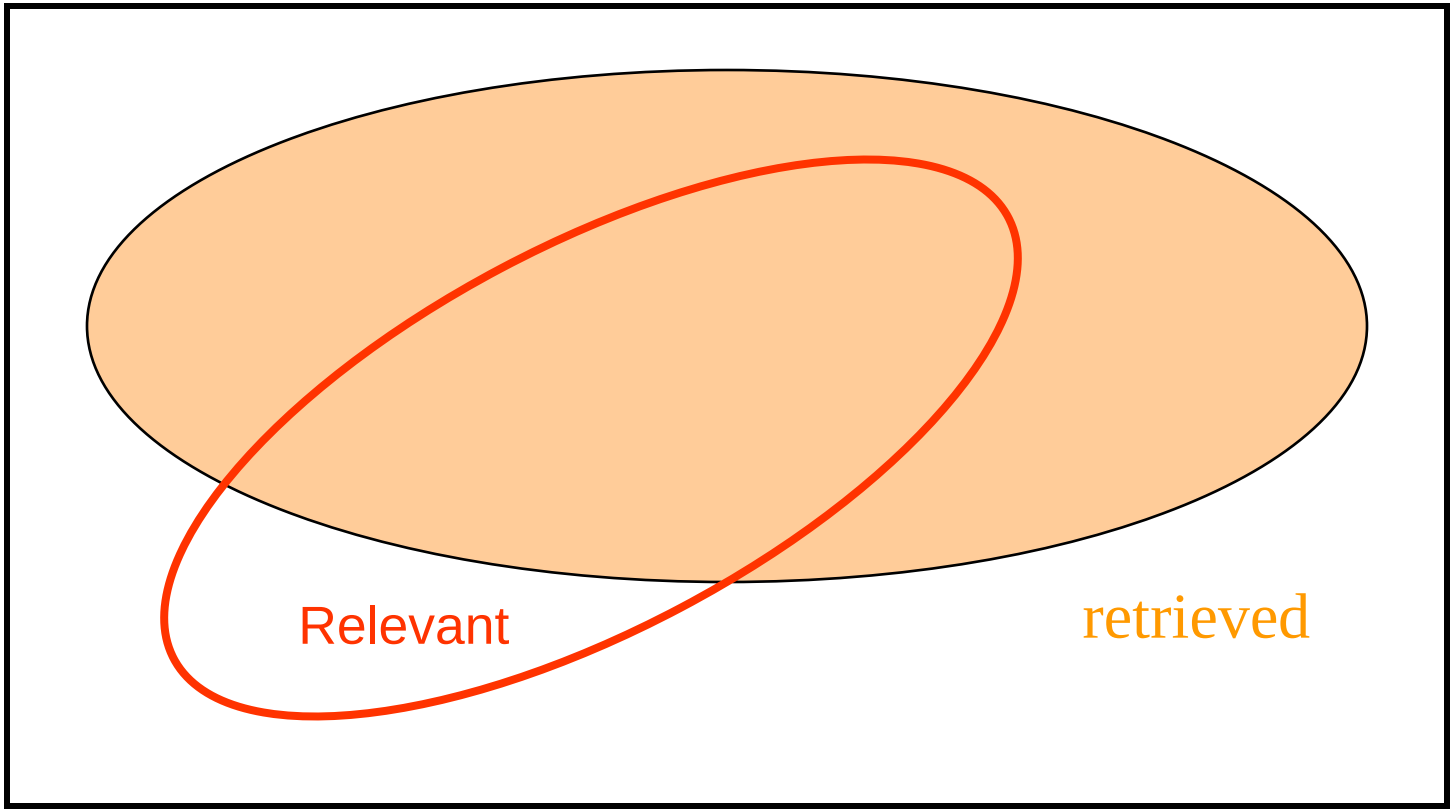
Very high precision, very low recall



Slide from Giles

Retrieved vs. Relevant Documents

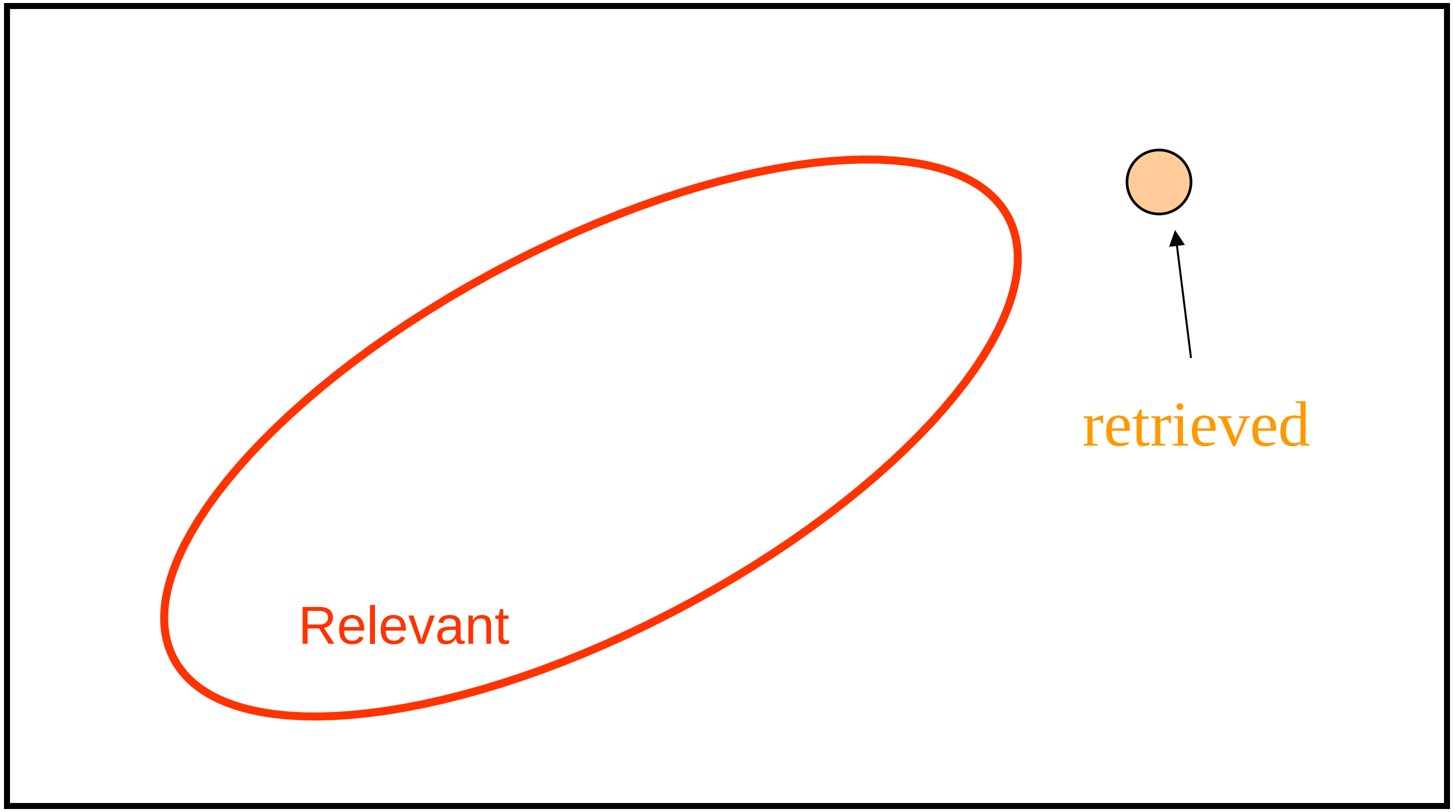
High recall, but low precision



Slide from Giles

Retrieved vs. Relevant Documents

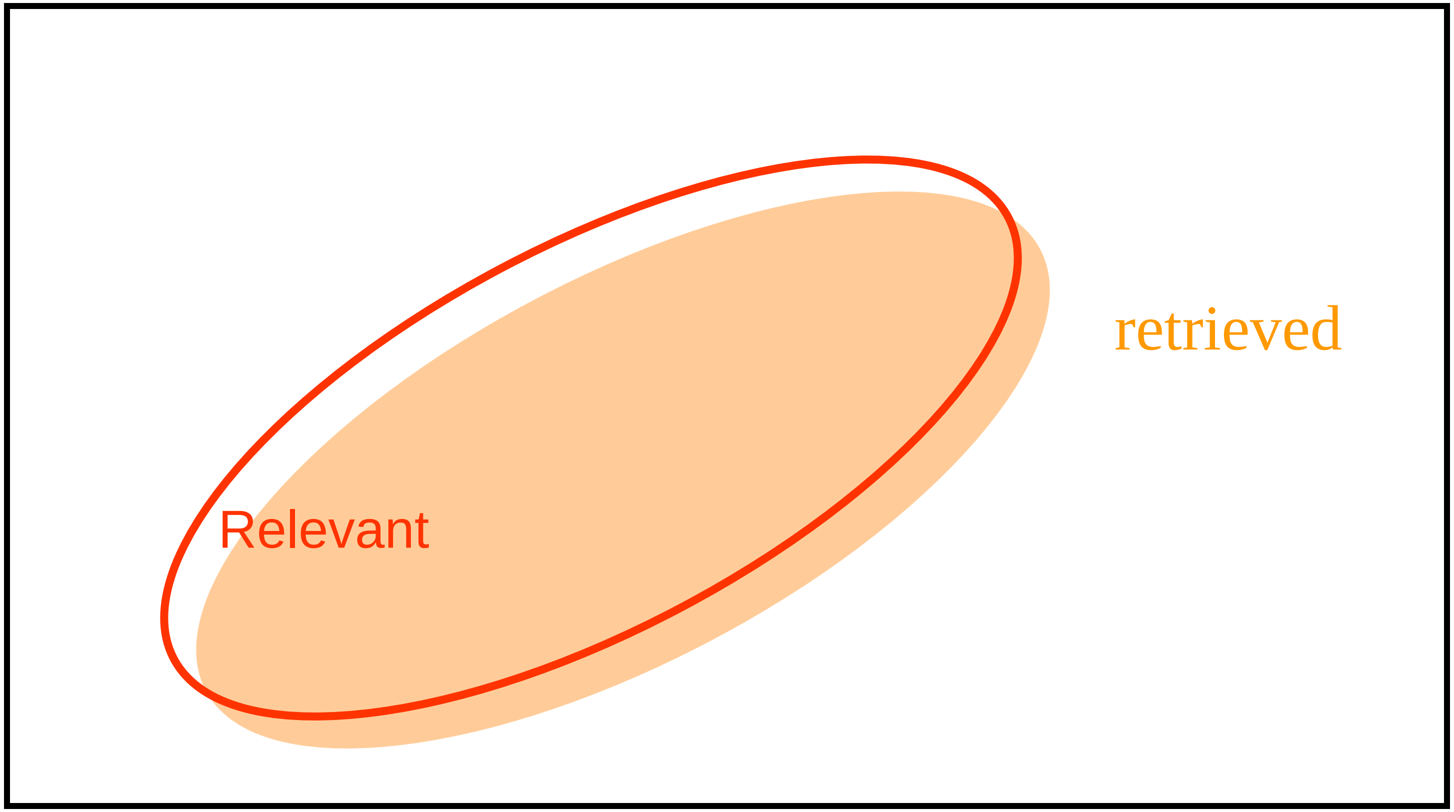
Very low precision, very low recall (0 for both)



Slide from Giles

Retrieved vs. Relevant Documents

High precision, high recall (at last!)



Slide from Giles

Why Precision and Recall?

Get as much of what we want while at the same time getting as little junk as possible.

Recall is the percentage of relevant documents returned compared to everything that is available!

Precision is the percentage of relevant documents compared to what is returned!

The desired trade-off between precision and recall is specific to the scenario we are in

Slide modified from Giles

Relation to Contingency Table

	Doc is Relevant	Doc is NOT relevant
Doc is retrieved	a	b
Doc is NOT retrieved	c	d

- Accuracy: (a+d) / (a+b+c+d)
- Precision: a/(a+b)
- Recall: a/(a+c)
- Why don't we use Accuracy for IR?
  - (Assuming a large collection)
    - Most docs aren't relevant
    - Most docs aren't retrieved
    - Inflates the accuracy value

Slide from Giles

CMU Seminars task

- Given an email about a seminar
- Annotate
  - Speaker
  - Start time
  - End time
  - Location

CMU Seminars - Example

<0.24.4.93.20.59.10.jgc+@NL.CS.CMU.EDU (Jaime Carbonell).0>
Type: cmu.cs.proj.mt
Topic: <speaker>Nagao</speaker> Talk
Dates: 26-Apr-93
Time: <stime>10:00</stime> - <etime>11:00 AM</etime>
PostedBy: jgc+ on 24-Apr-93 at 20:59 from NL.CS.CMU.EDU (Jaime Carbonell)

Abstract:

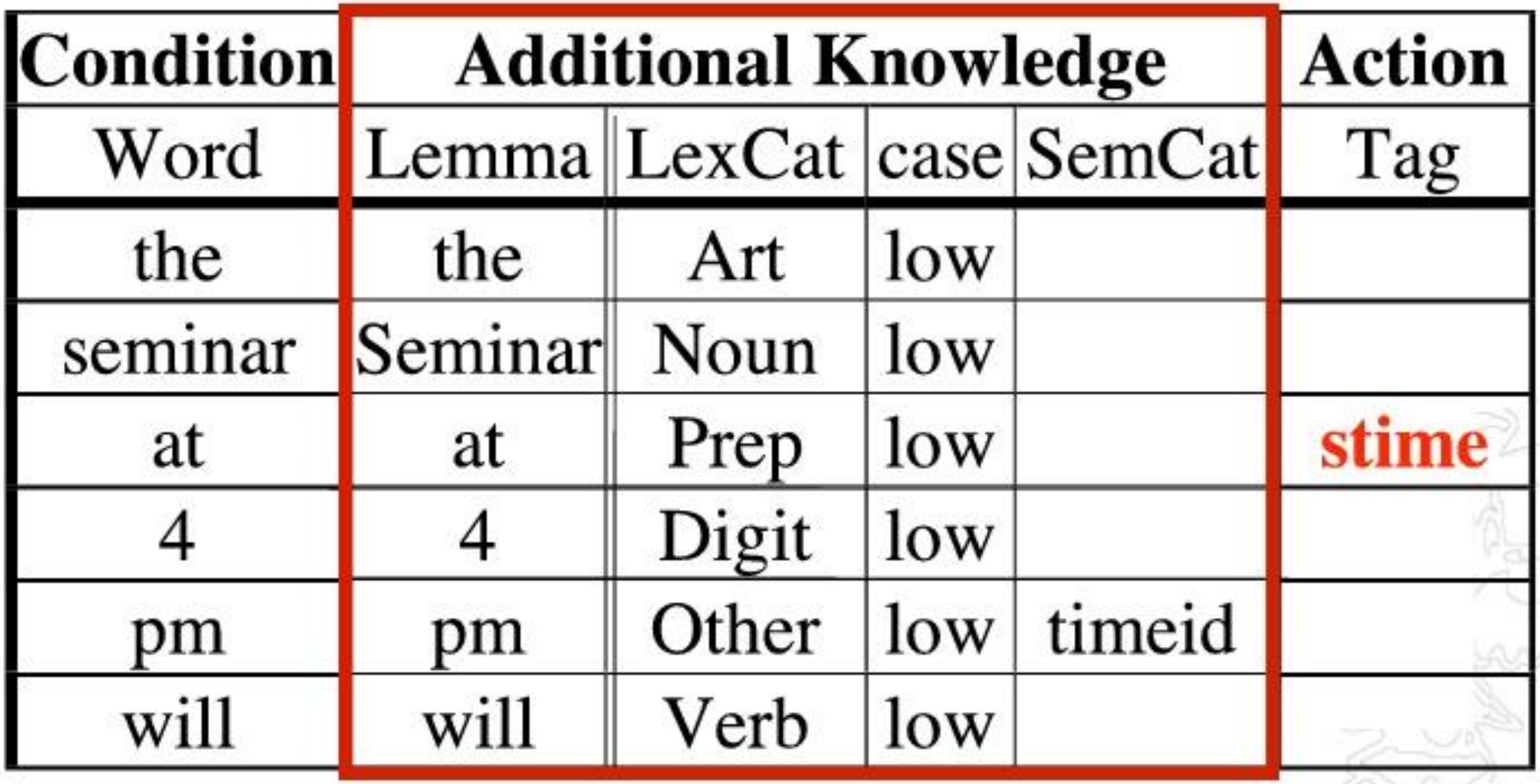
<paragraph><sentence>This Monday, 4/26, <speaker>Prof. Makoto Nagao</speaker> will give a seminar in the <location>CMT red conference room</location> <stime>10</stime>-<etime>11am</etime> on recent MT research results</sentence>.</paragraph>

Creating Rules

- Suppose we observe "the seminar at <stime>4 pm</stime> will [...]" in a training document
- The processed representation will have access to the words and to additional knowledge
- We can create a very specific rule for <stime>
  - And then generalize this by dropping constraints (as discussed previously)

Example

the seminar at <time> 4 pm will

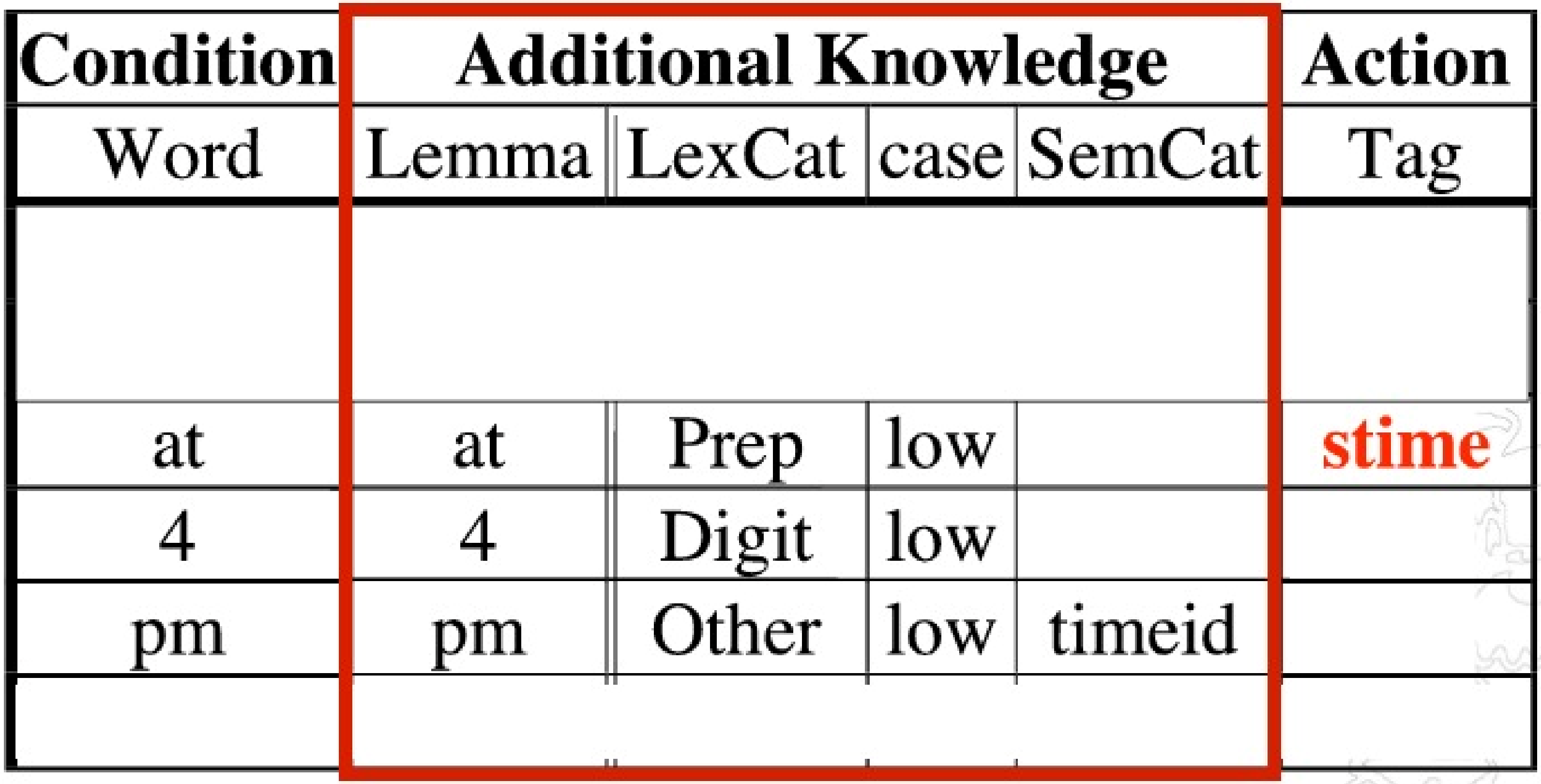


Condition	Additional Knowledge				Action
Word	Lemma	LexCat	case	SemCat	Tag
the	the	Art	low		
seminar	Seminar	Noun	low		
at	at	Prep	low		stime
4	4	Digit	low		
pm	pm	Other	low	timeid	
will	will	Verb	low		

Fabio Ciravegna:
Adaptive Information Extraction from Text by Rule Induction and Generalisation
in Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001), Seattle, August 2001.

Example

the seminar at <time> 4 pm will



Condition	Additional Knowledge				Action
Word	Lemma	LexCat	case	SemCat	Tag
at	at	Prep	low		stime
4	4	Digit	low		
pm	pm	Other	low	timeid	

Fabio Ciravegna:
Adaptive Information Extraction from Text by Rule Induction and Generalisation
in Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001), Seattle, August 2001.

Example

the seminar at <time> 4 pm will

Condition	Additional Knowledge				Action
Word	Lemma	LexCat	case	SemCat	Tag
	at				stime
		Digit			
				timeid	

Fabio Ciravegna:
Adaptive Information Extraction from Text by Rule Induction and Generalisation
in Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001), Seattle, August 2001.

- For each rule, we look for:
  - Support (training examples that match this pattern)
  - Conflicts (training examples that match this pattern with no annotation, or a different annotation)
- Suppose we see:

"tomorrow at <stime>9 am</stime>"

  - The rule in our example applies!
  - If there are no conflicts, we have a more general rule
- Overall: we try to take the most general rules which don't have conflicts

Returning to Evaluation

- This time, evaluation specifically for IE

Importance of Evaluation in IE

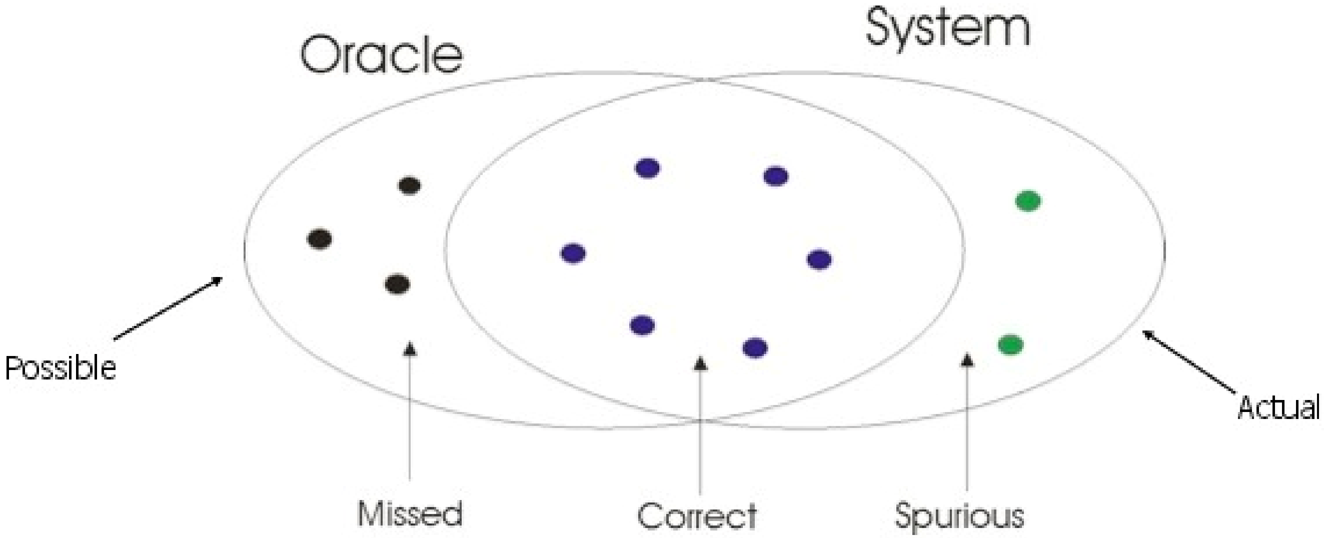
- IE was born from a series of competitive evaluations organised by DARPA in the US
  - MUC Conferences, 1989-1998
    - IE as a departure from IR but using the same types of measures of accuracy
    - The idea was to understand what worked and what not in text analysis
      - Finding a way to compare IE systems and approaches in a controlled way
- Evaluation is in IE's DNA
  - Publishing IE papers without evaluation is not considered acceptable

Organising Evaluation

- You will need:
  - An annotated training corpus
    - That you will use to develop rules or to train a machine learning algorithm
  - A result scorer
    - A tool that automatically computes accuracy of the system against an annotated corpus
    - E.g. The MUC Scorer
  - An annotated test corpus
    - To be used blindly to test results
      - Please note that run on test corpus should be a one off test
        - Test corpus is not be used to fine tuning accuracy in any way
        - E.g. By looking at the results and changing your rules or by tuning the learning parameters

The Rationale Behind

- **Precision**: how correct is the average answer provided by the system
- **Recall**: how many (correct) pieces of information are retrieved by the system
- **F-measure**: allows comparative evaluations



Evaluation Measures

$$\text{Recall} = \frac{\text{CORRECT} + (\text{PARTIAL} * 0.5)}{\text{POSSIBLE}}$$

$$\text{Precision} = \frac{\text{CORRECT} + (\text{PARTIAL} * 0.5)}{\text{ACTUAL}}$$

$$F(\beta) = \frac{(\beta^2 + 1) * \text{PREC} * \text{REC}}{\beta^2 * \text{PREC} + \text{REC}}$$

F-Measure is to be used to compare systems

In all evaluations all the three measures must be published

False Negative in CMU Seminars

- Gold standard test set:

  Starting from <stime>11 am</stime>

- System marks nothing:

  Starting from 11 am

- False negative (which measure does this hurt?)

False Positive in CMU Seminars

- Gold standard test set:

  ... Followed by lunch at 11:30 am , and meetings

- System marks:

  ... at <stime>11:30 am</stime>

- False positive (which measure does this hurt?)

Mislabeled in CMU Seminars

- Gold standard test set:

  at a different time - <stime>6 pm</stime>

- System marks:

  ... - <etime>6 pm</etime>

- What sort of error do we have here?
- Which measures are affected?
- Note that this is different from Information Retrieval!

Partial Matches in CMU Seminars

- Gold standard test set:

... at <stime>5 pm</stime>

- System marks:

... at <stime>5</stime> pm

- Then I get a partial match (worth 0.5)
- Also different from Information Retrieval

Issues in Evaluation

Issues Affecting Evaluation

- The Algorithm
- The feature set used
- The leniency in assessing results
    - the availability of standard annotated corpora do not guarantee that the experiments performed with different approaches and algorithms proposed in the literature can be reliably compared
        - Data problems
        - Problems of experimental design
        - Problems of presentation

Alberto Lavelli, Mary E Califf, Fabio Ciravegna, Dayne Freitag, Claudio Giuliano, Nicholas Kushmerick, Lorenza Romano, and Neil Ireson:
Evaluation of Machine Learning-based Information Extraction Algorithms: Criticisms and Recommendations, Language Resources and Evaluation, Volume 42, Issue 4 (December 2008).

Leniency in Evaluation

- Data Problems
  - Errors in data, branching corpora, templates Vs markup
- Experimental design
  - Training/Test Set selection
    - e.g. 50/50 Vs 80/20
  - Tokenization
  - How to count matches (see below)

Alberto Lavelli, Mary E Califf, Fabio Ciravegna, Dayne Freitag, Claudio Giuliano, Nicholas Kushmerick, Lorenza Romano, and Neil Ireson:
Evaluation of Machine Learning-based Information Extraction Algorithms: Criticisms and Recommendations, Language Resources and Evaluation, Volume 42, Issue 4 (December 2008).

Issues in Evaluation

- Fragment evaluation:
  - How leniently should inexact identification of filler boundaries be assessed?
- Counting multiple matches:
  - When a learner predicts multiple fillers for an entity, how should they be counted?
- Filler variation:
  - When text fragments having distinct surface forms refer to the same underlying entity, how should they be counted?

Alberto Lavelli, Mary E Califf, Fabio Ciravegna, Dayne Freitag, Claudio Giuliano, Nicholas Kushmerick, Lorenza Romano, and Neil Ireson: Evaluation of Machine Learning-based Information Extraction Algorithms: Criticisms and Recommendations, Language Resources and Evaluation, Volume 42, Issue 4 (December 2008).

- Evaluation is a critical issue where there is still much work to be done
- But before we can evaluate, we need a **gold standard**
- Training IE systems
  - Critical component for "learning" statistical classifiers
  - The more data, the better the classifier
- Can also be used for developing a handcrafted NER system
  - Constant rescoring and coverage checks are very helpful
- Necessary in both cases for **evaluation**

The University Of Sheffield.

Annotating Documents to IE Train Systems

Can we really ask people to annotate documents?

Most slides are from Ziqi Zhang, University of Sheffield

Do People Like Annotating?

- No, they hate it
  - They will try not to do it or do it quickly
- It is time and energy consuming
  - It is not their job
    - Unless they are professional annotators
  - They are not rewarded for it
- It is tiring
- It is error prone
- But most of all: is it possible to annotate documents with sufficient accuracy to train an IE system?

The archaeotools Experience

- A project funded by AHRC/EPSRC/JISC in the UK. In collaboration with the University of York (Archaeology Department)
- Goal:
  - Building an e-archaeology application to allow archaeologists to discover, share, and analyse datasets and legacy publications
- Role of IE: To identify in several collections of documents:
  - Pacenames: around 2,000 in corpus
    - Yorkshire, Cambridge, The London Tower, Baker Street, St. Paul, Church road.
  - Subjects: around 10,000
    - Roman pottery, spearhead, animal remains, church, courtyard, plates, vessel
  - Temporals: around 4,000
    - Roman, Saxon, AD1078, 300BC, 43 - 801AD, circa 1771, Victorian era, Bronze Age

http://nlp.shef.ac.uk/wig/research/ArchaeoTools.html

IE in Aracheotools

- Based on SVN
  - The TRex tool http://t-rex.sourceforge.net/
- Training based on corpora annotated by 5 expert archaeologists
  - training documents 42, length: up to several hundreds of pages
  - total documents to tag by machine learning: 967
  - total documents to tag by rules: 3991
- Annotation process was geared at high quality
  - Annotation instructions were clarified through several iterations
    - Our archaeologists colleagues, they clearly explained the task to annotators, went through examples with them
    - The IE experts went through several confusing examples with archaeologists to clarify their doubts
  - One senior researcher was appointed to make final decision in case of doubts from any annotators
  - Annotators were very motivated and the task was part of their job!!!

IE challenges – annotation quality

IAA F-measure – Inter-Annotator-Agreement F-measure, Hripcsak and Rothschild (2005).

		Annotator A	
		Positive	Negative
Annotator B	Positive	a	b
	Negative	c	d

- ✓ Treating A's annotations as gold standard, and B's as reference
- ✓ Precision of B = a/(a+b), Recall of B = a/(a+c)
- ✓ F-measure of B = 2a/(2a+b+c)
- ✓ Equivalent to the standard P, R, F metrics used for evaluating IE systems

Annotation quality (ctd.)

- IAA F-measure – Inter-Annotator-Agreement F-measure
  - ✓ Figures obtained from a shared corpus annotated by three different annotators

	Place name	Subject	Temporal
Lowest IAA between any two annotators	66.2	49	67.2
Highest IAA between any two annotators	*80*	*63*	*83.3*

Annotator Variability

- Differences in annotation are a significant problem
  - Only some people are good at annotation
  - Practice helps
- Even good annotators can have different understanding of the task
  - For instance, in doubt, annotate? Or not?
  - (~ precision/recall tradeoffs)
- Effect of using gold standard corpora that are not well annotated
  - Evaluations can return inaccurate results
  - Systems trained on inconsistent data can develop problems which are worse than if the training examples are eliminated
- Crowd-sourcing, which we will talk about later, has all of these same problems even more strongly!

Annotation Quality - Conclusions

- In general archaeology is a difficult domain, with many uncertainty and ambiguity even for humans
- Inconsistency between annotators generated noise that influences learning system
- Very careful evaluation of the quality of annotation must always be implemented
  - Aka possibility/ability for the annotators to perform good quality annotation
- Never ever suppose that humans are 100% correct
  - For complex tasks they may perform at 80% accuracy!!!!
    - Always ask users to annotate (at least partially) overlapping sets of documents
      - So to be able to check their agreement

- Slide sources
  - Some of the slides presented today were from C. Lee Giles, Penn State and Fabio Ciravegna, Sheffield

Conclusion

- Last two lectures
  - Rule-based NER
  - Learning rules for NER
  - Evaluation
  - Annotation

- Please read Sarawagi Chapter 3!

- Thank you for your attention!