Start Recording

- Today:
  - Evaluation

Reminders

- Project 4 is due Tonight at 11:45 PM.
- Writing assignment is due Tuesday, September 4.
- Project 5 is due Thursday, September 6.
- Course evaluations now available

Evaluation

- Contemporary research in computational linguistics is unacceptable if it is not accompanied by principled evaluation

> "An important recent development in NLP has been the use of much more rigorous standards for the evaluation of NLP systems"
>
> Manning and Schutze

- Quantitative measurement of results is what differentiates our field from armchair theorizing
- To be published, all research must:
  - establish a baseline, and
  - quantitatively show that it improves on the baseline
- To be better, you should:
  - evaluate *why* it's a good baseline
  - describe *how* your system performed better

slide adapted from Glenn Slaydon

Basic Evaluation

- "How well does the system work?"
- Possible domains for evaluation
  - Processing time of the system
    - (familiar from project 4)
  - Space usage of the system
  - Human satisfaction
  - Correctness of results

slide adapted from Glenn Slaydon

Example

- You are building a system which automatically provides short, human-readable summaries of a set of documents on a given topic
- Your system picks sentences from the documents based on word co-occurrences, and presents these sentences as the summary
- We want to evaluate the “quality” of the results
- One choice in such a system is whether you should use stemming when determining the word co-occurrences
- Let’s briefly examine stemming

slide adapted from Glenn Slaydon

Stemming

- Morphological suffixes used can make our data more sparse
- In content-analysis tasks (IE, IR, summarization), we may only care about ‘stems,’ because they carry the “content” of the lemma

example:

He doesn’t like to shop.

She shops at the mall.

I went shopping last week.

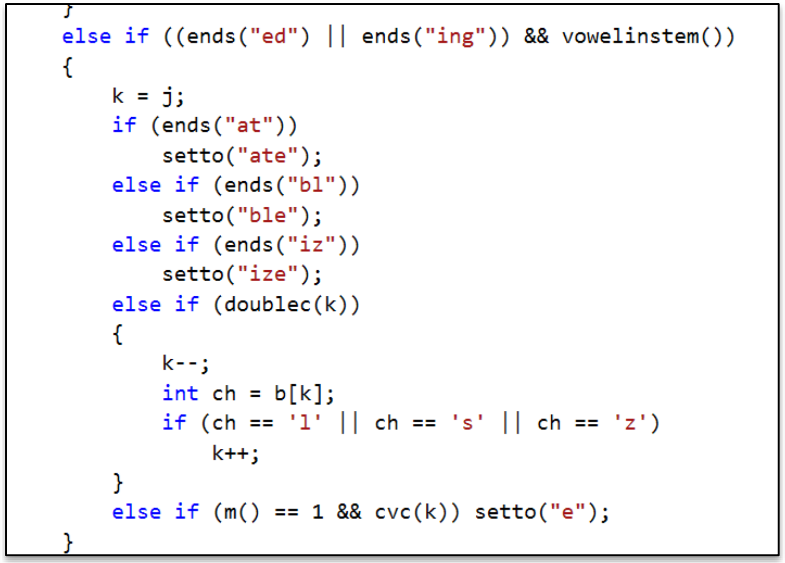
Ben shopped until he dropped.

Bill is quite an avid shopper.

slide adapted from Glenn Slaydon

Porter stemmer

- One well known stemming algorithm for English is the Porter stemmer
  - M. F. Porter. 1980. *An algorithm for suffix stripping*. Program, 14(3):130–137.
- It is a heuristic which contains lots of code like this:



```
}
else if ((ends("ed") || ends("ing")) && vowelinstem())
{
    k = j;
    if (ends("at"))
        setto("ate");
    else if (ends("bl"))
        setto("ble");
    else if (ends("iz"))
        setto("ize");
    else if (doublec(k))
    {
        k--;
        int ch = b[k];
        if (ch == 'l' || ch == 's' || ch == 'z')
            k++;
    }
    else if (m() == 1 && cvc(k)) setto("e");
}
```

slide adapted from Glenn Slaydon

Porter stemmer

```
Two medical experts testifying Wednesday in the doping
trial of a former East German sports doctor said the female
swimmers they examined showed health damage linked to
performance-enhancing drugs, including liver damage and
excessive facial hair.
```

```
two medic expert testifi wednesdai in the dope trial of a
former east german sport doctor said the femal swimmer thei
examin show health damag link to performance-enhanc drug
includ liver damag and excess facial hair
```

slide adapted from Glenn Slaydon

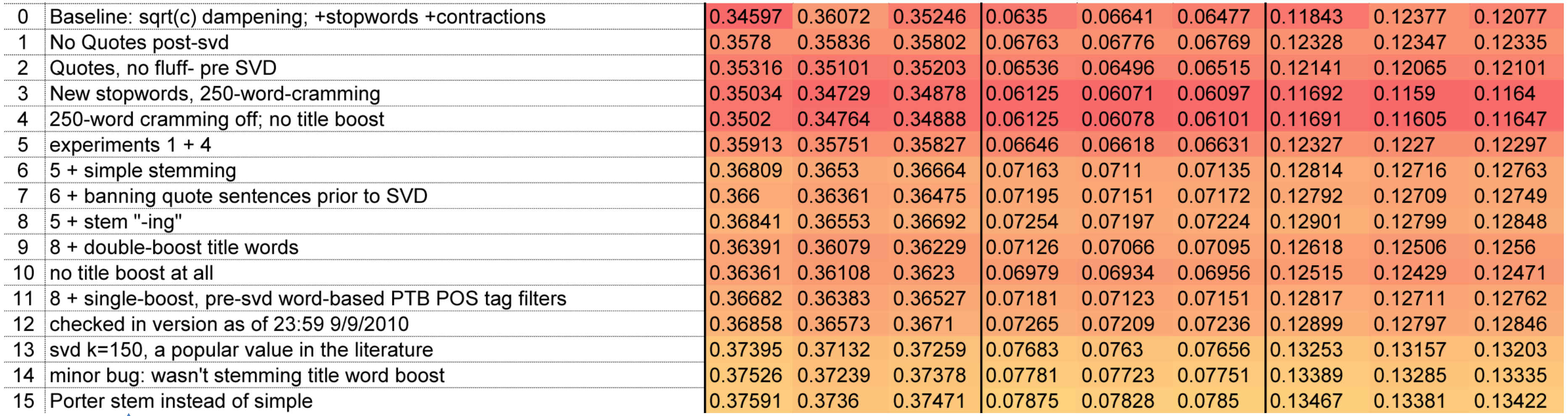
Back to our example

- Looking at the output of the Porter stemmer makes linguists cringe
- For the document summarizer system:

> “We elected not to [illegible] stemmer because we were getting goo[illegible]s with our system already, and we di[illegible]ns the linguistic data that we had ca[illegible]nto gibberish.”

slide adapted from Glenn Slaydon

Evaluate!

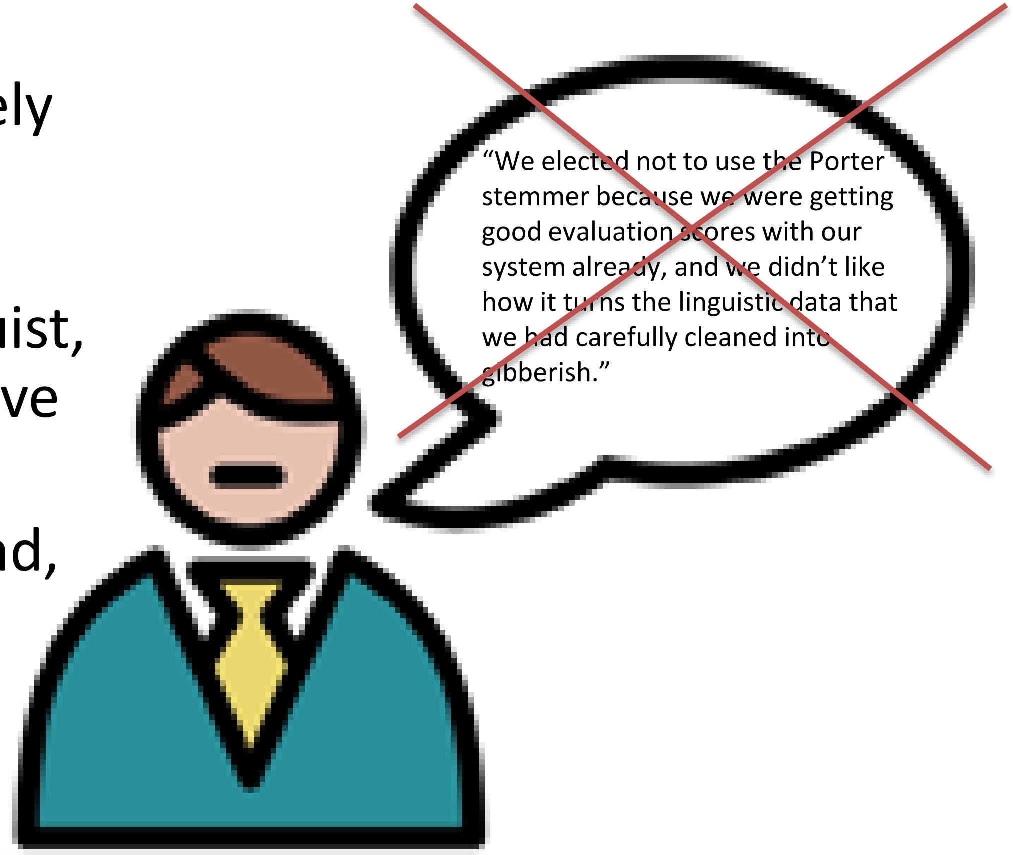


		ROUGE-1			ROUGE-2			ROUGE-SU4		
		min	max	avg	min	max	avg	min	max	avg
0	Baseline: sqrt(c) dampening; +stopwords +contractions	0.34597	0.36072	0.35246	0.0635	0.06641	0.06477	0.11843	0.12377	0.12077
1	No Quotes post-svd	0.3578	0.35836	0.35802	0.06763	0.06776	0.06769	0.12328	0.12347	0.12335
2	Quotes, no fluff- pre SVD	0.35316	0.35101	0.35203	0.06536	0.06496	0.06515	0.12141	0.12065	0.12101
3	New stopwords, 250-word-cramming	0.35034	0.34729	0.34878	0.06125	0.06071	0.06097	0.11692	0.1159	0.1164
4	250-word cramming off; no title boost	0.3502	0.34764	0.34888	0.06125	0.06078	0.06101	0.11691	0.11605	0.11647
5	experiments 1 + 4	0.35913	0.35751	0.35827	0.06646	0.06618	0.06631	0.12327	0.1227	0.12297
6	5 + simple stemming	0.36809	0.3653	0.36664	0.07163	0.0711	0.07135	0.12814	0.12716	0.12763
7	6 + banning quote sentences prior to SVD	0.366	0.36361	0.36475	0.07195	0.07151	0.07172	0.12792	0.12709	0.12749
8	5 + stem "-ing"	0.36841	0.36553	0.36692	0.07254	0.07197	0.07224	0.12901	0.12799	0.12848
9	8 + double-boost title words	0.36391	0.36079	0.36229	0.07126	0.07066	0.07095	0.12618	0.12506	0.1256
10	no title boost at all	0.36361	0.36108	0.3623	0.06979	0.06934	0.06956	0.12515	0.12429	0.12471
11	8 + single-boost, pre-svd word-based PTB POS tag filters	0.36682	0.36383	0.36527	0.07181	0.07123	0.07151	0.12817	0.12711	0.12762
12	checked in version as of 23:59 9/9/2010	0.36858	0.36573	0.3671	0.07265	0.07209	0.07236	0.12899	0.12797	0.12846
13	svd k=150, a popular value in the literature	0.37395	0.37132	0.37259	0.07683	0.0763	0.07656	0.13253	0.13157	0.13203
14	minor bug: wasn't stemming title word boost	0.37526	0.37239	0.37378	0.07781	0.07723	0.07751	0.13389	0.13285	0.13335
15	Porter stem instead of simple	0.37591	0.3736	0.37471	0.07875	0.07828	0.0785	0.13467	0.13381	0.13422

slide adapted from Glenn Slaydon

Now who’s cringing?

- Although the crude approach of the Porter stemmer seems intuitively ugly, *it improves our evaluation metric*
- As a computational linguist, you must remain objective
- If you have a legitimate linguistic intuition in mind, test it!
- Evaluate, evaluate, evaluate


~~“We elected not to use the Porter stemmer because we were getting good evaluation scores with our system already, and we didn’t like how it turns the linguistic data that we had carefully cleaned into gibberish.”~~

slide adapted from Glenn Slaydon

Now who’s cringing?

- As a professional computational linguist, it is this type of statement that should immediately set off your alarm bells
- Always ask:
  - What is the BASELINE?
  - What is the RESULT?
  - How was it EVALUATED?

~~“We elected not to use the Porter stemmer because we were getting good evaluation scores with our system already, and we didn’t like how it turns the linguistic data that we had carefully cleaned into gibberish.”~~

slide adapted from Glenn Slaydon

Stemming and Performance

- Does stemming help in IR, IE, and document summarization?
- Harman 1991 indicated that it hurt as much as it helped
  - D. Harman (1991) How effective is suffixing. In *Journal of the American Society for Information Science.* 42(7) 7-15.
- Krovetz 1993 shows that it does help
  - R. Krovetz (1993) Viewing morphology as an inference process. In *Proc. 16th ACM SIGR R&D IR* 191-202
  - Porter-like algorithms work well with smaller documents
  - Krovetz proposes that stemming loses information
  - Derivational morphemes tell us something that helps identify word senses; stemming them loses this

slide adapted from Glenn Slaydon

Evaluating a stemmer

- In the summarization example, the stemmer is a small part of the whole system
- We used an end-to-end measurement (ROUGE scores) to evaluate the impact of using stemming
- How would you evaluate the “performance” of stemming by itself?

slide adapted from Glenn Slaydon

“Correct” stemming

- This would be difficult, because there’s no “correct” stemming of a word
- The best stemmer is a hash function that conflates all words with the same linguistic stem: the actual stemmed token doesn’t matter

Stemmer #7

eat	04xBrLt
ate	04xBrLt
eating	04xBrLt
eats	04xBrLt

Better than…

Porter Stemmer

eat	eat
ate	at
eating	eat
eats	eat

slide adapted from Glenn Slaydon

Therefore

- Results are sensitive to the specific application.
- Rule of thumb: when there’s an implementation decision to be made:
  - Evaluate the alternatives
  - Document your results and choice
- Some choices may require going back and re-evaluating earlier decisions

slide adapted from Glenn Slaydon

Accuracy

- In order to evaluate your system, you need to know what is correct or desired
- A set of data which is labeled with the correct or desired result is called a gold standard
- If the system you are evaluating is a function that maps one input to one output, then you can evaluate accuracy
  - Correct: matches the gold standard
  - Incorrect: otherwise

$$accuracy = \frac{number\ correct}{number\ of\ results}$$

slide adapted from Glenn Slaydon

Accuracy: example

- With your Naïve Bayesian classifier for language identification, say we are only interested in the single best language it selects for a given input sentence

"مساء الخير" → Arabic ✓

"Bon soir!" → French ✓

"Good evening!" → Spanish ✗

Accuracy:
.67
67%

slide adapted from Glenn Slaydon

Error

- You can also measure error. This is the proportion of items that you got wrong

$$error = \frac{number\ wrong}{number\ of\ results}$$

“مساء الخير” → Arabic ✓

“Bon soir!” → French ✓

“Good evening!” → Spanish ✗

Error:
.33
33%

slide adapted from Glenn Slaydon

Evaluating classifiers

- Many NLP problems involve classifying things
- For these problems, there is a more nuanced way to evaluate performance
- Also, note that many NLP problems can be re-stated as classification problems
- Before we talk about evaluating classification results, let’s see how to re-state a problem as a classification problem

slide adapted from Glenn Slaydon

Example: Named Entity Recognition

- A string of text, pick out which substrings are named entities.

Bohr founded the Institute of Theoretical Physics at the University of Copenhagen, now known as the Niels Bohr Institute, which opened in 1920. Bohr mentored and collaborated with physicists including Hans Kramers, Oskar Klein, George de Hevesy, and Werner Heisenberg.

- How can this be reframed as a classification problem?

slide adapted from Glenn Slaydon

Example: Named Entity Recognition

- Each word-terminating character (space, punctuation, start/end of line) either begins (BNE 🟥) or ends (ENE ▮) of a named entity, or does nothing.

🟥Bohr▮founded the🟥Institute of Theoretical Physics▮at the🟥University of Copenhagen▮, now known as the🟥Niels Bohr Institute▮, which opened in 1920.🟥Bohr▮mentored and collaborated with physicists including🟥Hans Kramers▮,🟥Oskar Klein▮,🟥George de Hevesy▮, and🟥Werner Heisenberg▮.

slide adapted from Glenn Slaydon

Evaluating Classifiers

- Classifiers divide items into different categories
- They “label” items, or put them into different sets
- For each label/category/set, you can say:
  - There are items which are selected
  - There are items which are not selected
- The gold standard tells you which items should have been selected, i.e. “correct”

slide adapted from Glenn Slaydon

Precision and Recall

- Precision and Recall are set-based measures
- They evaluate the quality of some set membership, based on a reference set membership

slide adapted from Glenn Slaydon

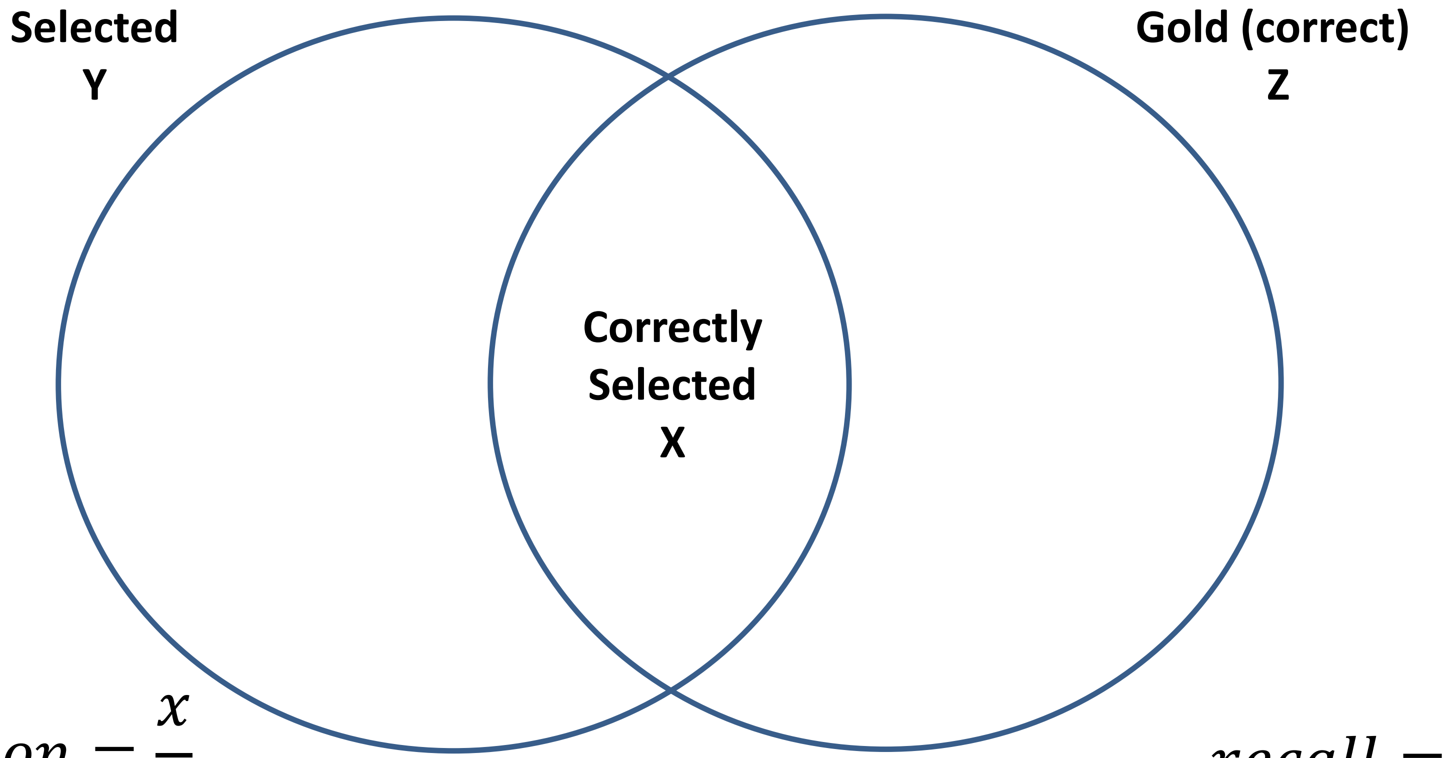
Precision and Recall

Precision: what proportion of the *selected* items are in the gold set?

Recall: how many *items from the gold set* got selected?

slide adapted from Glenn Slaydon

Precision / Recall



$$precision = \frac{x}{y}$$

$$recall = \frac{x}{z}$$

$$\Omega = y \cup z$$

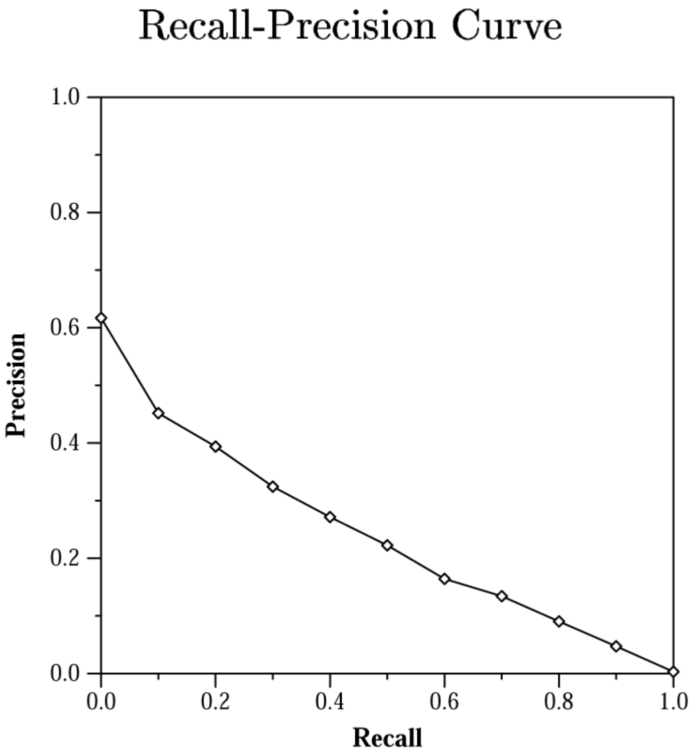
slide adapted from Glenn Slaydon

Precision-recall trade-off

- Usually, precision and recall can be traded for each other by changing your system parameters

The higher the proportion of correct items you require in the selected set (high precision), the fewer of the total correct items you will select (low recall)

If you can tolerate a higher proportion of incorrect items in the selected set (low precision), you will capture more of the total correct items (high recall)

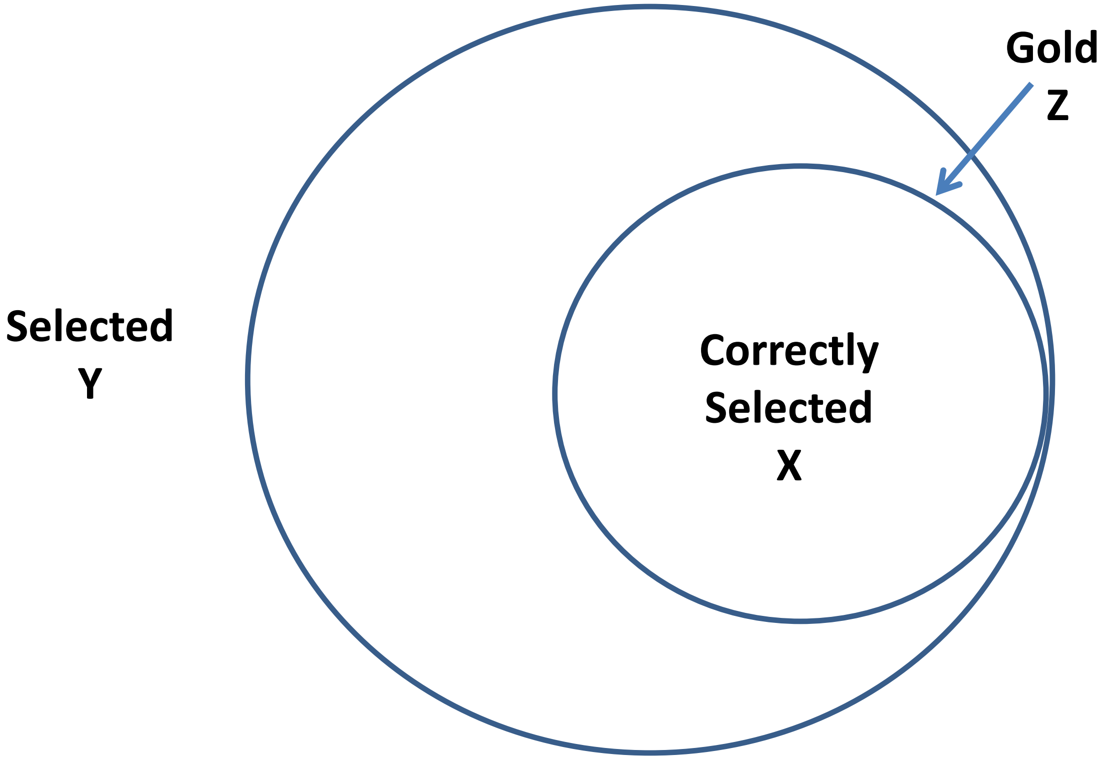


slide adapted from Glenn Slaydon

Precision-recall trade-off

- It’s easy to get recall of 1.0. Why?

- Return all the documents



$$recall = \frac{x}{z}$$

slide adapted from Glenn Slaydon

Summary

		Gold standard	
		True	False
Your Result	True	true positive tp	false positive fp type I error
	False	false negative fn type II error	true negative tn

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \qquad precision = \frac{tp}{tp + fp}$$

$$error = \frac{fp + fn}{tp + tn + fp + fn} \qquad recall = \frac{tp}{tp + fn} \qquad \left(fallout = \frac{fp}{fp + tn}\right)$$

slide adapted from Glenn Slaydon

F-measure

- a way of reporting precision and recall together in one number
- based on van Rijsbergen (1979):

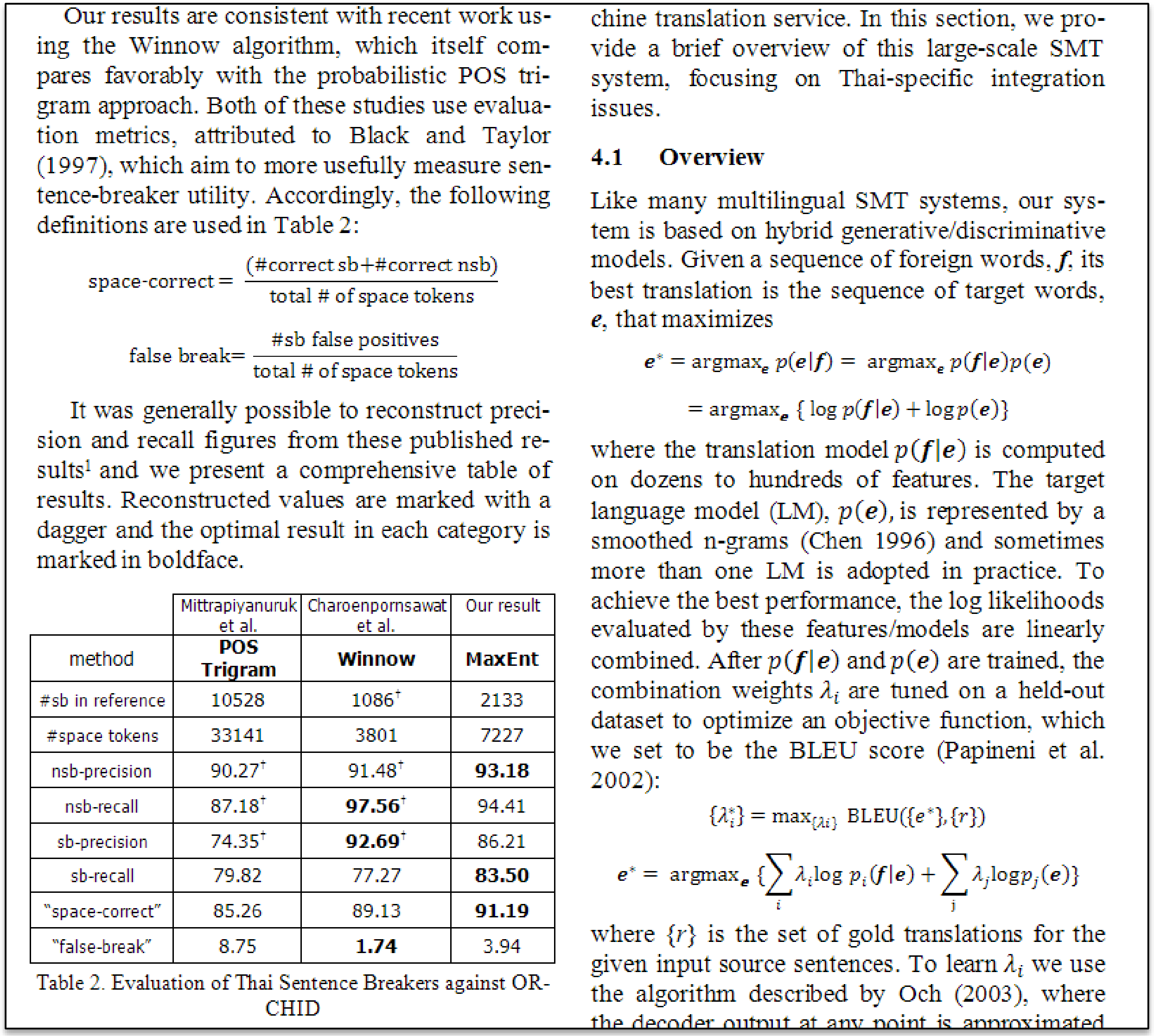
$$F = \frac{2PR}{R + P}$$

- This is a simple average; P and R are each weighted by .5
- Other versions of F scores use different weights for P versus R

slide adapted from Glenn Slaydon

Reading the literature

- Refereed journal papers will have Results and Evaluation sections
- Evaluation is often shown in a table versus previous work
- If skimming papers, skip to the results and evaluation section for a summary of the work



Our results are consistent with recent work using the Winnow algorithm, which itself compares favorably with the probabilistic POS trigram approach. Both of these studies use evaluation metrics, attributed to Black and Taylor (1997), which aim to more usefully measure sentence-breaker utility. Accordingly, the following definitions are used in Table 2:

$$\text{space-correct} = \frac{(\text{\#correct sb} + \text{\#correct nsb})}{\text{total \# of space tokens}}$$

$$\text{false break} = \frac{\text{\#sb false positives}}{\text{total \# of space tokens}}$$

It was generally possible to reconstruct precision and recall figures from these published results[1] and we present a comprehensive table of results. Reconstructed values are marked with a dagger and the optimal result in each category is marked in boldface.

	Mittrapiyanuruk et al.	Charoenpornsawat et al.	Our result
method	**POS Trigram**	**Winnow**	**MaxEnt**
#sb in reference	10528	1086†	2133
#space tokens	33141	3801	7227
nsb-precision	90.27†	91.48†	**93.18**
nsb-recall	87.18†	**97.56**†	94.41
sb-precision	74.35†	**92.69**†	86.21
sb-recall	79.82	77.27	**83.50**
"space-correct"	85.26	89.13	**91.19**
"false-break"	8.75	**1.74**	3.94

Table 2. Evaluation of Thai Sentence Breakers against ORCHID

chine translation service. In this section, we provide a brief overview of this large-scale SMT system, focusing on Thai-specific integration issues.

4.1 Overview

Like many multilingual SMT systems, our system is based on hybrid generative/discriminative models. Given a sequence of foreign words, $\boldsymbol{f}$, its best translation is the sequence of target words, $\boldsymbol{e}$, that maximizes

$$\boldsymbol{e}^* = \operatorname{argmax}_{\boldsymbol{e}} p(\boldsymbol{e}|\boldsymbol{f}) = \operatorname{argmax}_{\boldsymbol{e}} p(\boldsymbol{f}|\boldsymbol{e})p(\boldsymbol{e})$$

$$= \operatorname{argmax}_{\boldsymbol{e}} \{\log p(\boldsymbol{f}|\boldsymbol{e}) + \log p(\boldsymbol{e})\}$$

where the translation model $p(\boldsymbol{f}|\boldsymbol{e})$ is computed on dozens to hundreds of features. The target language model (LM), $p(\boldsymbol{e})$, is represented by a smoothed n-grams (Chen 1996) and sometimes more than one LM is adopted in practice. To achieve the best performance, the log likelihoods evaluated by these features/models are linearly combined. After $p(\boldsymbol{f}|\boldsymbol{e})$ and $p(\boldsymbol{e})$ are trained, the combination weights λ_i are tuned on a held-out dataset to optimize an objective function, which we set to be the BLEU score (Papineni et al. 2002):

$$\{\lambda_i^*\} = \max_{\{\lambda i\}} \text{BLEU}(\{e^*\},\{r\})$$

$$\boldsymbol{e}^* = \operatorname{argmax}_{\boldsymbol{e}} \{\sum_i \lambda_i \log p_i(\boldsymbol{f}|\boldsymbol{e}) + \sum_j \lambda_j \log p_j(\boldsymbol{e})\}$$

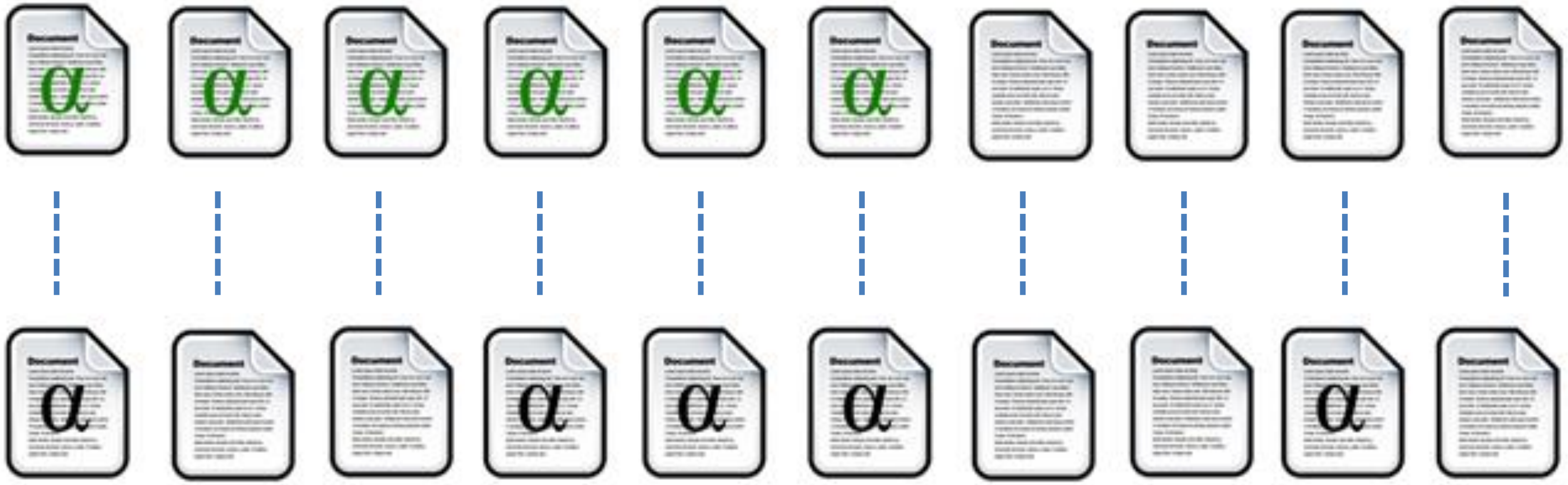
where $\{r\}$ is the set of gold translations for the given input source sentences. To learn λ_i we use the algorithm described by Och (2003), where the decoder output at any point is approximated

http://research.microsoft.com/apps/pubs/default.aspx?id=130868

slide adapted from Glenn Slaydon

In-class quiz #1

This is the gold standard for biomedical documents which mention the IL−2R α−promoter. The result of our classifier is shown below.



What is the accuracy? .70

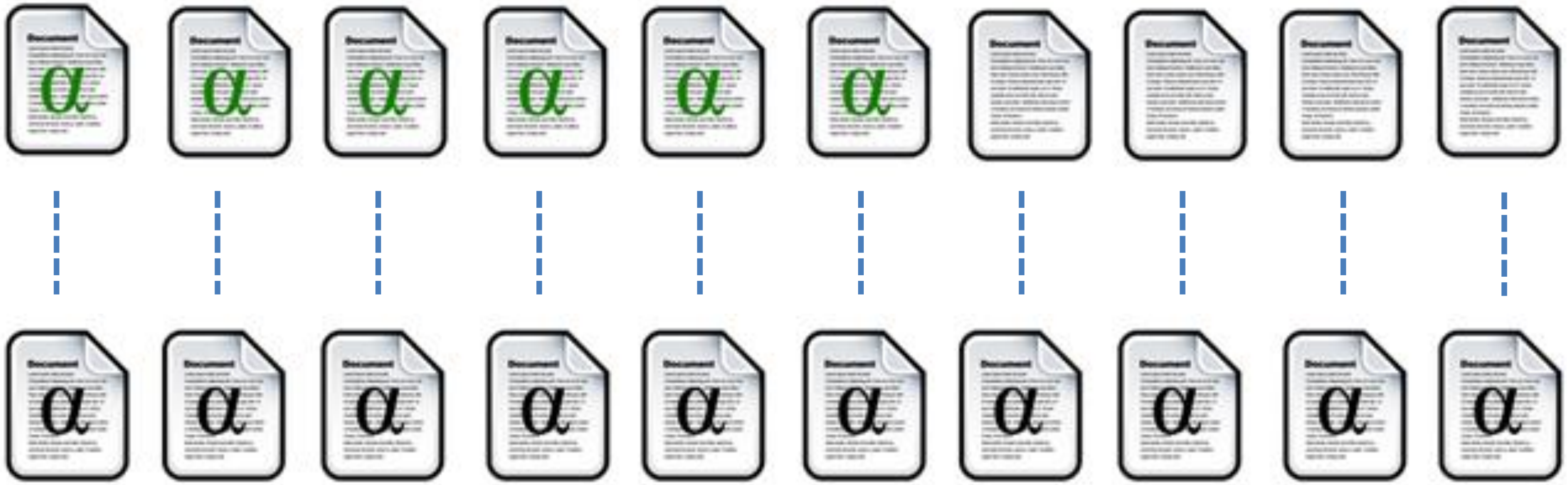
What is the precision? .80

What is the recall? .66

slide adapted from Glenn Slaydon

In-class quiz #2

This is the gold standard for biomedical documents which mention the IL−2R α−promoter. The result of our classifier is shown below.



What is the accuracy? .60

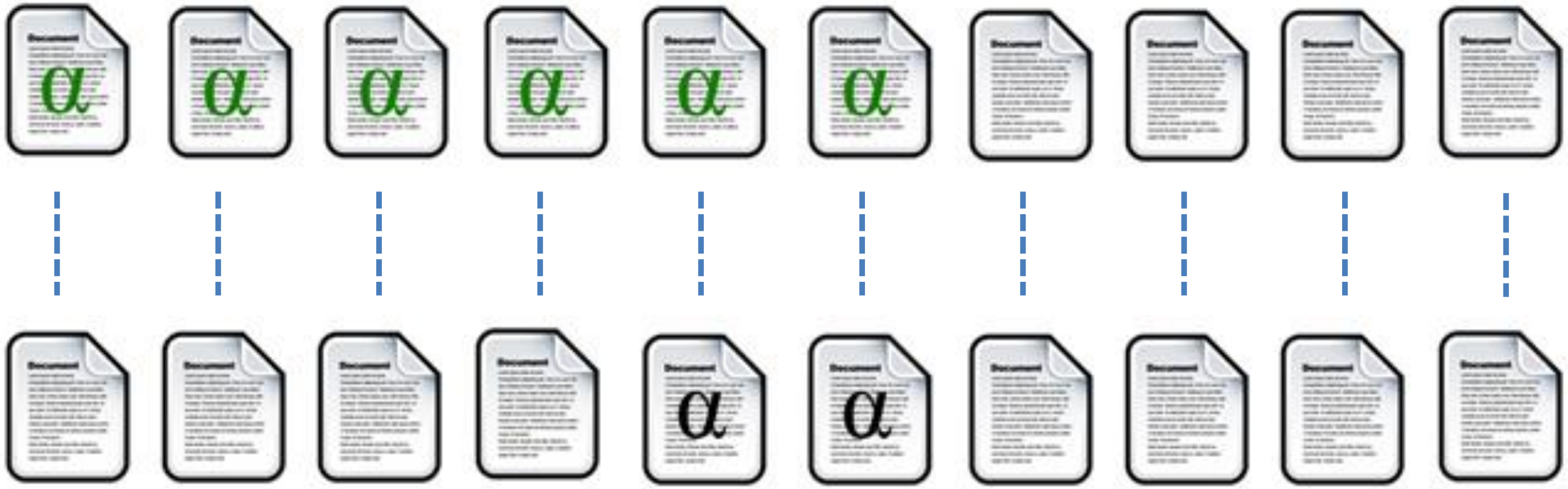
What is the precision? .60

What is the recall? 1.0

slide adapted from Glenn Slaydon

In-class quiz #3

This is the gold standard for biomedical documents which mention the IL−2R α−promoter. The result of our classifier is shown below.



What is the accuracy? .60

What is the precision? 1.0

What is the recall? .33

slide adapted from Glenn Slaydon

Example: Information Retrieval

- You need to have high recall to be useful for downstream tasks
- Good recall but poor precision means that you find all the instances, but pick a lot of false examples
  - The true positives will get lost in the noise of false positives
- Good precision but poor recall means that all the instances you find are accurate, but you miss a lot of them
  - How will you find the information you're looking for?
- Information retrieval prefers recall over precision

slide adapted from Glenn Slaydon

Planning for evaluation

- Design evaluation strategy at the start of your research
  - What will be measured?
  - What is the baseline?
    - Is there a state of the art available?
    - Don’t spend a lot of time building a complex system that does worse than “pick the most common example” (e.g., pick the most common POS)
  - What is the gold standard (reference)?
  - What is the measurement heuristic?
    - Maximizing recall, precision (on what random variable)?

slide adapted from Glenn Slaydon

Measure/metric/indicator

- a measure:
  - a figure, extent, or amount obtained by measuring
- a metric (distance)
  - the degree to which a system possesses a given attribute, or
  - a combination of two or more measures
- an indicator
  - the amount of deviation from a baseline state (+0.7, -1.3, …)

slide adapted from Glenn Slaydon

Standardized evaluation

- When evaluating complex systems, it’s helpful to use the same measure that was used to measure comparable (competitive) systems
- Machine translation:
  - BLEU, NIST, METEOR
- Document summarization:
  - ROUGE-1, ROUGE-2, ROUGE-SU4, Pyramid

slide adapted from Glenn Slaydon