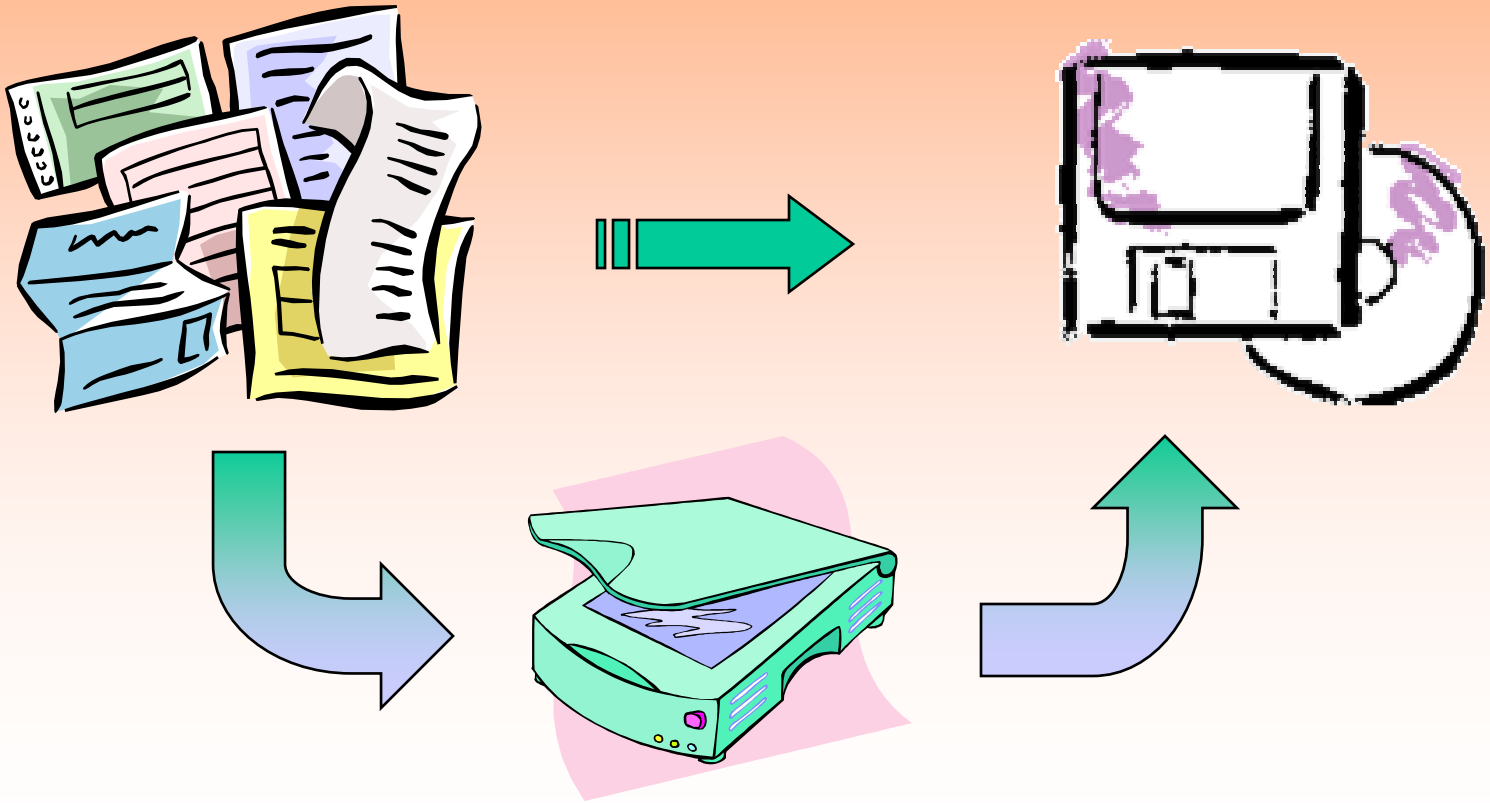
A Voting System for Automatic Correction of OCR Output

Tomi Klein **Miri Kopel**

Bar Ilan University

Introduction

OCR = Optical Character Recognition



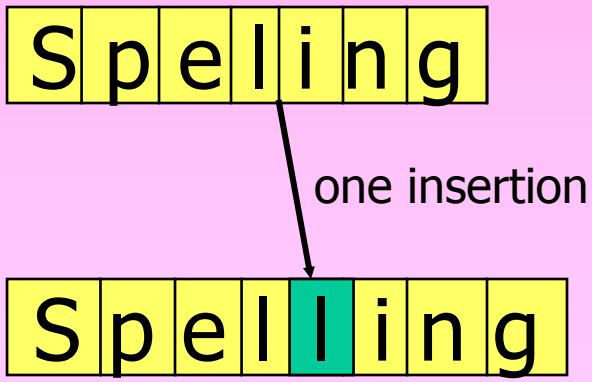
Problems:

- **Even the most accurate OCR devices don't give 100% accuracy.**
- **Manual correction is too expensive**

Known Techniques for Spelling Correction

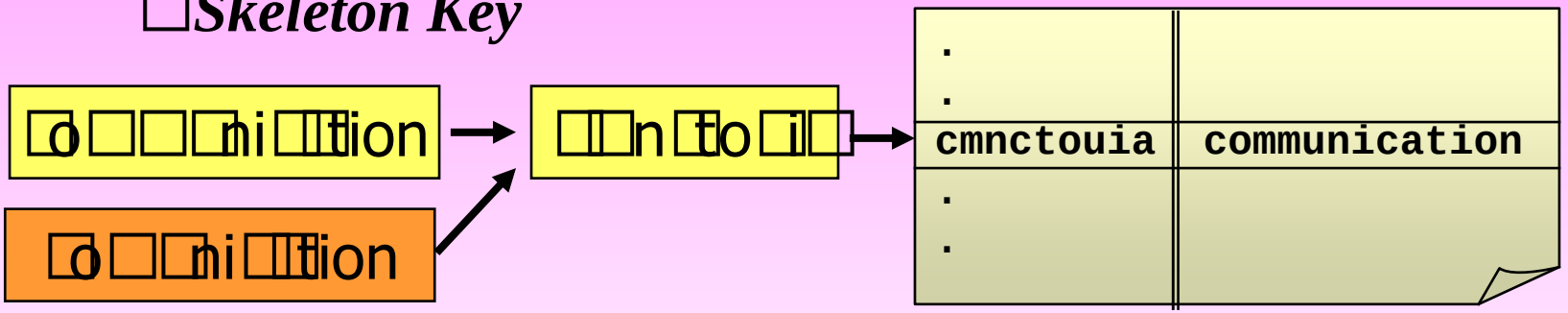
Edit Distance:

The minimum number of editing operations (i.e., insertion, deletion and substitution of letters) required to transform one string to another.

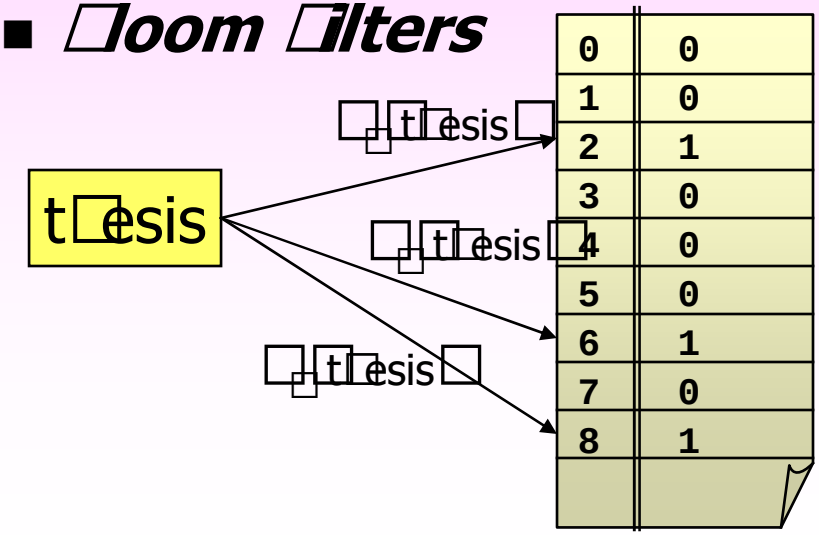


Known Techniques for Spelling Correction (cont...)

- **Hashing:**
  - *Skeleton Key*



  - *Bloom Filters*



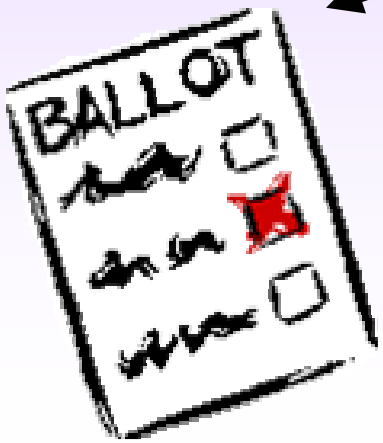
So, what is the problem ?

- **Most of the techniques are relevant only for typing errors.**

- **Designed for isolated words.**

- **We are interested in a fully automatic system.**

The Algorithm



- □□□□□□□□□s its o□n t□pi□□□ errors□□i□□i□□er ro□ e□□□ ot□er□
- □or e□□□□□□e□i□e □□i□□□
- ***con□usion matrix***□
- □si□□□r□□ter □□□□r□□□□

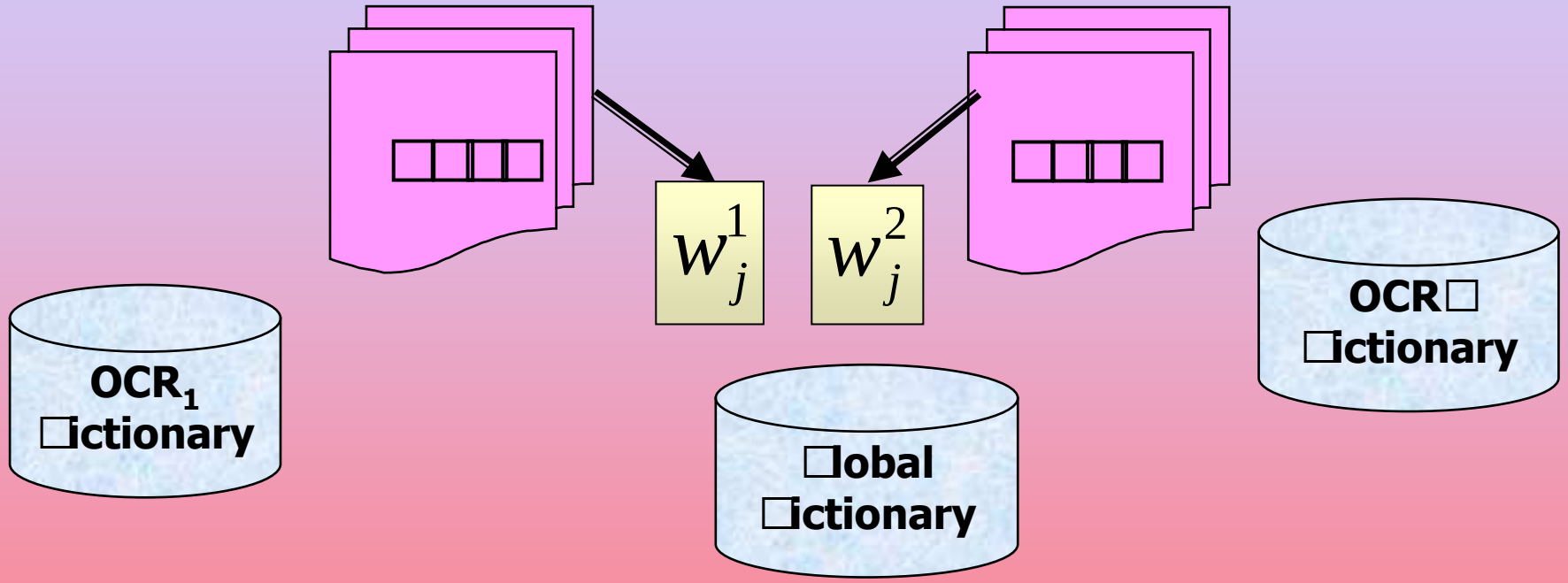
The Algorithm (cont...)

Initialization Phase

- Build the dictionary.
- Create the confusion matrices.
- Construct the *n*-gram and word-gram tables.
- Build the documents sets.

The Algorithm (cont...)

□etection □nd Correction Phase



□□□□e □or□s □re i□enti□□□□
□□□□e □or□s □is□t□□□□□□t onl□one o□t□e□ is □□□i□□
□□□□e □or□s □is□t□□□□n□none o□t□e□ is □□□i□□
□□□□e □or□s □is□t□□□□n□□ot□re □□□i□□

The Algorithm (cont...)

- If words are identical:
  Accept the word.

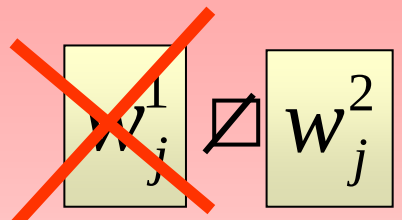
$$w_j^1 = w_j^2$$

- If the words are not identical:
  - Dictionaries
  - Use character errors
  - Confusion matrix
  - Context

$$w_j^1 \neq w_j^2$$

The Algorithm (cont...)

If the words mismatch:



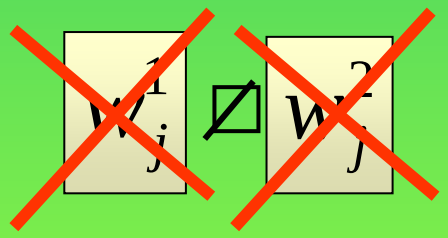
- If only one of them is valid Then:
  Accept the valid one.

$$dictionary(w_j^i) = 0.6 \cdot freq(w_j^i, local\_dictionary) + 0.4 \cdot freq(w_j^i, global\_dictionary)$$

The Algorithm (cont...)

- Candidate generation

  Generate candidates for both w_j^1 and w_j^2 and then prune the least one



$$Candidates = \text{Generate\_Candidates}(w_j^1) \cup \text{Generate\_Candidates}(w_j^2)$$

The Algorithm (cont...)

□□r□□eters □or r□n□ing t□e □□n□i□□ties□

□□□i□tion□ries□ □□□□it □ist□n□e□

□□□or□□□r□□□ □□□on□□sion □□tri□□

$$is\_close = \begin{cases} 1 & \text{if } edit\_distance(w^i_j, w^i_{j_k}) = 1. \\ 0 & \text{otherwise.} \end{cases}$$

$$\begin{aligned} mark(w^i_{j_k}) = 0.6 \cdot (error\_freq(w^i_j, w^i_{j_k}, OCR^i)) + \\ 0.4 \cdot (word\_gram(w^i_{j-1}, w^i_{j_k}, w^i_{j+1}) + \\ dictionary(w^i_{j_k}) + \\ is\_close(w^i_j, w^i_{j_k})) \end{aligned}$$

The Algorithm (cont...)

- **If both words are valid:**

$$context(w_j^i) = freq(w_j^i, local\_dictionary) + word\_gram(w_{j-1}^i, w_j^i, w_{j+1}^i).$$

$$mark(w_j^i) = accuracy(OCR^i) \cdot [0.6 \cdot context(w_j^i) + 0.4 \cdot freq(w_j^i, global\_dictionary)]$$

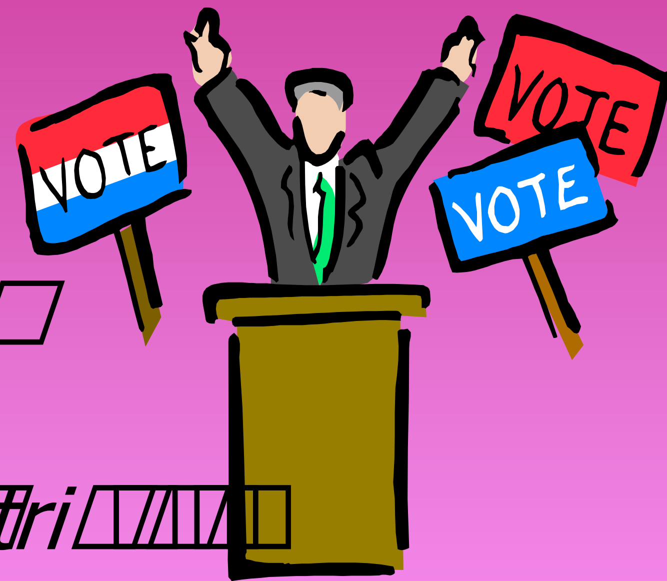
The Algorithm (cont...)

Generate Candidates(w)

[illegible] $_n$

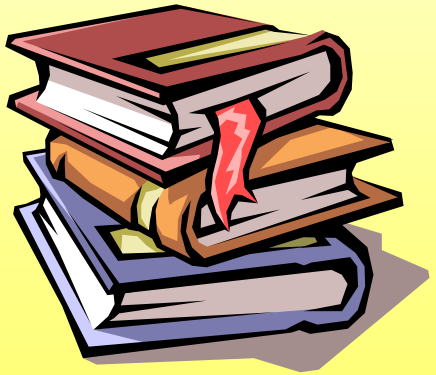
↓

[illegible]or [illegible]l [illegible] $_i$ [illegible] $_i$ [illegible] [illegible]o[illegible]



- [illegible]e [illegible] *on[illegible]sion[illegible]tri[illegible]*
- [illegible]t[illegible]ere [illegible]re r[illegible]les [illegible] [illegible]ppl[illegible]t[illegible]e[illegible]
- [illegible]t[illegible]e gener[illegible]te[illegible]or[illegible]is [illegible]li[illegible]
  [illegible]it to t[illegible]e [illegible]n[illegible]i[illegible]tes set[illegible]

The Experiments



The Environment of the Experiments

The confusion matrix for English text

Source String	*Error String*	*OCR 1*	*OCR 2*
i	1	**1.58%**	**0.48%**
0	o	**3.54%**	**0.77%**
da	d	**0.75%**	**0.10%**
2	z	**0.30%**	**0.72%**
f	l	**0.15%**	**0.72%**
i	1	**1.58%**	**0.48%**
e	g	**1.18%**	**0.10%**

Examples of successful corrections

English and French texts

Original Word	*OCR 1*	*OCR 2*	*Accepted Word*
going	going	goring	going
survivors	surveyors	survivors	survivors
we’re	we’te	we’e	we’re
thankfully	thankfvlly	thankfiily	thankfully
school	schooi	sciiool	school
neighborhood	ne~hborhood	nei&iborho~	neighborhood
precisent	pr,cisent	preci~ent	precisent
details	d,tails	detaHs	details

Examples of erroneous decisions

Errors can be generated for several reasons:

1. **We accept an incorrect word directly from the OCR (for e.g. when both OCRs generate the same mistake).**
2. **We choose the wrong word between two valid alternatives.**
3. **Only one of the words is in the dictionary but it is not the correct one.**
4. **We (corrected) the input word which was correct to an incorrect one.**
5. **We couldn't find a proper candidate for the misspelled input word.**

Original Word	*OCR 1*	*OCR 2*	*Accepted Word*	*Error type*
5,600	5,6oo	5,6oo	5,6oo	1
Ian	lan	lan	lan	1
class	dass	dass	dass	1
100	100	too	too	2
true	true	toe	toe	2
little	lithe	lisle	lithe	2
We	the	We	the	2
circles	circ.les	circus	circus	3
says	savs	sas	san	4
2001	2ooi	zooi	zoo	4

Analysis of the Results

Model 1:
OCR1, dictionary with no post processing.

Model 2:
OCR2, dictionary with no post processing.

Model □:
OCR1 □
post processing

Model □:
OCR2 □
post processing

Analysis of the Results (cont…)

Model □:
OCR 1 □ OCR 2:
si□ple dictionary
loo□□p.

Model □:
OCR 1 □ OCR 2:
dictionary loo□□p
with
pro□a□ilities.

Model □:
OCR 1 □ OCR 2:
□□ll post
processing.

Analysis of the Results (cont…)

		Error Rate
1	**OCR 1 – no post-processing**	14.0%
2	**OCR 2 – no post-processing**	12.1%
3	**OCR 1 – dictionary + candidates generation**	8.5%
4	**OCR 2 – dictionary + candidates generation**	8.0%
5	**Both – comparison + simple dictionary lookup**	7.2%
6	**Both – comparison + dictionary's frequencies**	5.1%
7	**Both – full correction system**	3.6%

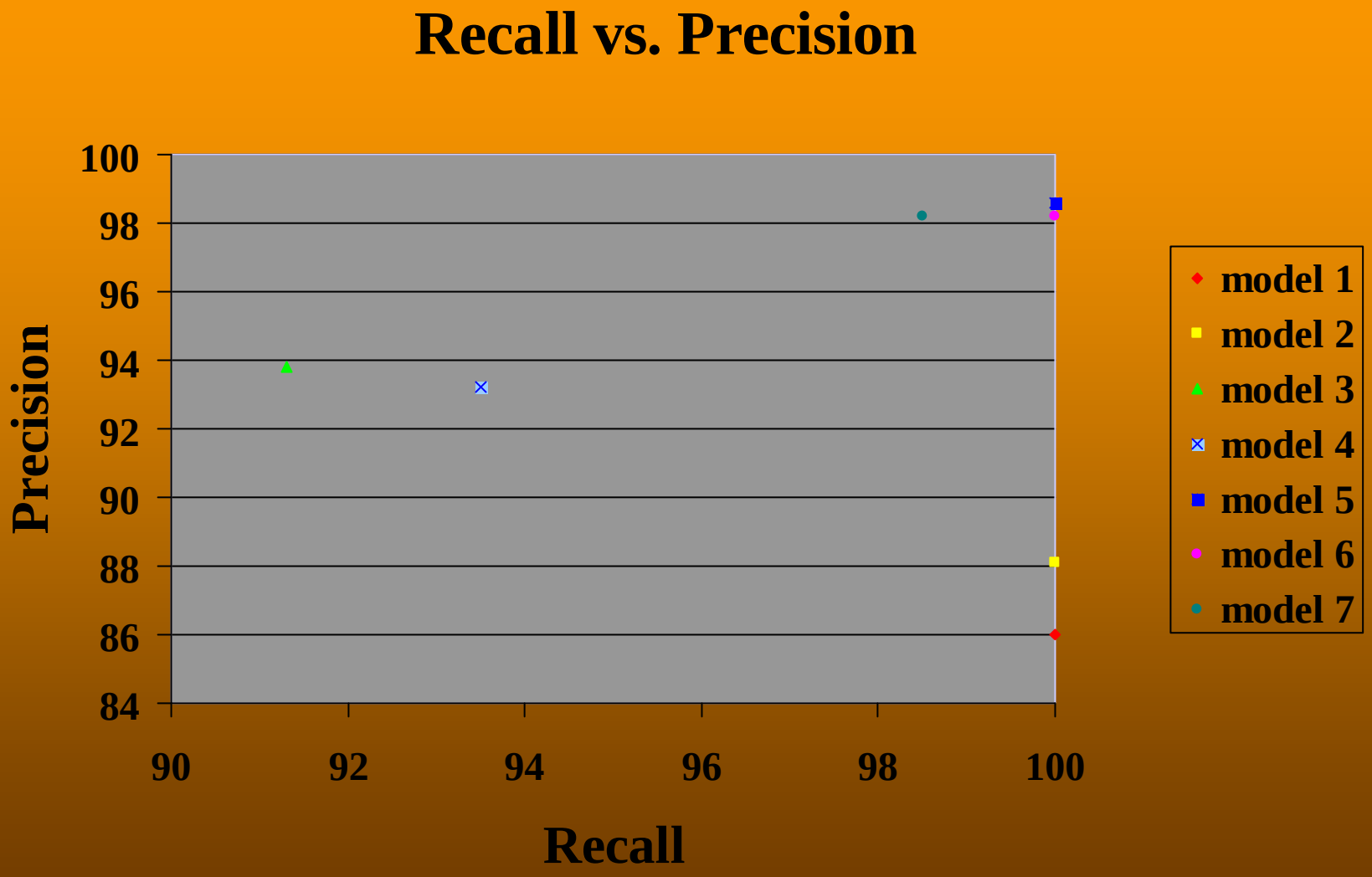
Analysis of the Results (cont…)

	Accept words from OCR as is	Do not accept the OCR word, and try to suggest candidates
Correct word written to output	**(A) true positive**	**(B) true negative**
Incorrect word written to output	**(C) false positive**	**(D) false negative**

$$\text{Recall} = \frac{A}{A+B} = \frac{\#\text{ correct words accepted directly from OCR}}{\#\text{ correct words accepted}}$$

$$\text{Precision} = \frac{A}{A+C} = \frac{\#\text{ correct words accepted directly from OCR}}{\#\text{ total words accepted directly from OCR}}$$

Analysis of the Results (cont…)



Further Work

- More OCR devices.
- Context: NLP techniques, word class
- Specifications for certain language.