Performance Evaluation for Text Processing of Noisy Inputs

Daniel Lopresti

Computer Science & Engineering
Lehigh University
Bethlehem, PA 18015, USA
lopresti@cse.lehigh.edu

Motivation

Earlier attempt to study impact of errors from optical character recognition (OCR) on automatic summarization:

- use shallow language understanding (tokenization, PoS tagging),
- apply past statistics (word frequencies, sentence positions),
- look for cue phrases (e.g., "In conclusion ..."),
- extract key sentences (or phrases or paragraphs) for summary.

Situation might arise when processing large quantities of scanned documents (e.g., information extraction, digital libraries).

But many of same basic text processing steps apply elsewhere.

"Summarizing Noisy Documents," H. Jing, D. Lopresti, and C. Shih, *Proceedings of the Symposium on Document Image Understanding Technology*, April 2003, Greenbelt, MD, pp. 111-119.

Cut-and-paste Text Summarization, H. Jing, Ph.D. Thesis, Dept. of Computer Science, Columbia University, 2001.

On Clean Input ...

Kingdom To Sign Nuclear Non-proliferation Treaty

Saudi Arabia on Tuesday decided to sign the nuclear weapons non-proliferation treaty, a strong indication it will not seek nuclear warheads for intermediate-range missiles it recently acquired from China.

The official Saudi Press Agency reported that King Fahd made the decision during a Cabinet meeting in Riyadh, the Saudi capital.

The meeting was called in response to a recommendation by Prince Saud al-Faisal, the Saudi foreign minister, that the kingdom sign the international treaty against the spread of nuclear arms.

An account of the Cabinet discussions and decisions at the meeting, which ended before dawn, was issued by Information Minister Ali al-Shaer and distributed by the agency. The agency, monitored in Bahrain, did not elaborate.

It appeared the timing of the decision was designed primarily to reassure the United States that the kingdom will not try to arm its CSS-2 missiles with nuclear warheads. The decision also was viewed as an attempt to blunt Israel's allegations that the missiles constituted a threat to its safety.

Saudi Arabia, the Middle East petroleum giant and the world's largest exporter of crude oil, was reported to have recently acquired from Beijing an undisclosed number of CSS-2 missiles capable of reaching virtually any point in the Middle East, including Israel.

Israel has voiced fears the Saudis might be seeking to acquire nuclear warheads for the missiles and indicated it might deal a preemptive blow.

Ideal Summary (Human)

Saudi Arabia on Tuesday decided to sign the nuclear weapons non-proliferation treaty, a strong indication it will not seek nuclear warheads for intermediate-range missiles it recently acquired from China.

It appeared the timing of the decision was designed primarily to reassure the United States that the kingdom will not try to arm its CSS-2 missiles with nuclear warheads.

Automatic Summary

Saudi Arabia on Tuesday decided to sign the nuclear weapons non-proliferation treaty, a strong indication it will not seek nuclear warheads for intermediate-range missiles it recently acquired from China.

Saudi Arabia, the Middle East petroleum giant and the world's largest exporter of crude oil, was reported to have recently acquired from Beijing an undisclosed number of CSS-2 missiles capable of reaching virtually any point in the Middle East, including Israel.

On Noisy Input ...

Ideal Summary (from Original Document)

Fans of the Baltimore Orioles are laughing on the outside, but crying on the inside as they watch their team soar like a stone.

Wednesday's loss to the Milwaukee Brewers gave the Orioles the dubious distinction of being the first major league team in history to lose 14 games at the start of the season.

They continued their losing ways Thursday, falling 7-1 to the Brewers in Milwaukee.

But fans in Birdland say it will take more than a string of losses to kill the pride they have in the team that won the World Series five years ago and six pennants from 1966 to 1983.

Patty Waters, an administrative assistant in the Orioles' public relations office, said the telephones have been ringing off the hook as fans called to offer encouragement and suggestions.

Automatic Summary (from OCR of Light Photocopy)

Both wcar a ZIOVE for no apparent reason- That's one of the jokes makin- the rounds about the winless team that's also known as the Zer-O's.

Fans of the Baltimore Orioles are lau.@hina on the outside, but c@,ina on the inside as the)! watch their team soar like a stone.

-,Ogj NVednesday's loss to the Milwaukee Brewers gave the Orioles the dubious distinction of btina the first major]careful team in history to lose 1 4 aames at the start of the season, They continued their losina ways Thursday, fallinl, 7-1 to the Brewers

But fans in Birdland say it will take more than a strine of losses to kill the pride they have in the team that won he World Series five years a-o and six pennants from 1966 to 2983. think lhev'.-c the @a@d C,@,@s Slov,!;lns'l-i, 1-8, 2 sales representative for a beer company.

... I can set the 9@20] b?,us under m@,, eves," he said- "But T'm holdina up my Commitments Pattv Waters, an administrative assistant in the Orioles' public relations office, said the telephones have off ,he book- as fans called to offer encouragement and su--@stions.

"One lady wanted to hold a -,na,@s positive thinkin- seminar for the fans and the club," axis.

I (Jon't know if f@-71 the teacher was mal-i@., but the h-,"d,' (the p!aN,er-s) --'he@, -um and stick it on the -nd of their bats to help make contact, or put it in their E!lo-,,es so they might catch a ball, @: Nls.

Lots of things going on here: want to get to bottom of this.

Previous Attempt at Evaluation

Traditional approach to evaluating automatic summarization computes overlap (e.g., unigram, bigram) with human summaries.

To measure relative impact of OCR errors, compute overlap between automatic summaries based on noisy and clean text inputs.

To try to localize effects, we took closer look at individual stages:

- Classify OCR errors using string edit distance.
- Evaluate sentence boundary detection performance by comparing total number of sentences and average words / sentence.
- Evaluate PoS tagging errors by counting number of incomplete parse trees.

Last two measures are indirect – not satisfying. Hence this paper ...

Text Processing Stages: Functions

Processing Stage	*Intended Function*
Optical character recognition	Transcribe input bitmap into encoded text (hopefully accurately).
Sentence boundary detection	Break input into sentence-sized units, one per text line.
Tokenization	Break each sentence into word (or word-like) tokens delimited by white space.
Part-of-speech tagging	Takes tokenized text and attaches label to each token indicating its part-of-speech.

Text Processing Stages: Problems

Processing Stage	*Potential Problem(s)*
Optical character recognition	Current OCR is "brittle," errors made early-on propagate to later stages.
Sentence boundary detection	Missing or spurious sentence boundaries due to OCR errors on punctuation.
Tokenization	Missing or spurious tokens due to OCR errors on whitespace and punctuation.
Part-of-speech tagging	Bad PoS tags due to failed tokenization or OCR errors that alter orthographies.

Problems 1

Sentence boundary detection results for clean input:

CHAPTER 1 Loomings.

Call me Ishmael.

Some years ago--never mind how long precisely--having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world.

Results for noisy input (light photocopy):

' cH__ ' R l ' .

_omings.

, call me IshMael.

soMe ye_s ago--never mind how long , p,ec;sely__hav;ng _;€tle or no _oney in my purse, and nothing p_;,u__ to ;,terest Me on shore, I thoug_t I would sail _boUt a _;tt1e and see _e watery p_ or the world.

Note: 3 sentences vs. 4 sentences.

Problems 2

Tokenization results for clean input:

CHAPTER 1 Loomings .

Call me Ishmael .

Some years ago -- never mind how long precisely -- having little or no money in my purse , and nothing particular to interest me on shore , I thought I would sail about a little and see the watery part of the world .

Results for noisy input (light photocopy):

' cH__ ' R 1 ' .

_omings .

, call me IshMael .

soMe ye_s ago -- never mind how long , p , ec ; sely__hav ; ng _ ; €tle or no _oney in my purse , and nothing p_ ; , u__ to ; , terest Me on shore , I thoug_t I would sail _boUt a _ ; tt1e and see _e watery p_ or the world .

Problems 3

Part-of-speech tagging results for clean input:

CHAPTER_NNP 1_CD Loomings_NNS ._.

Call_VB me_PRP Ishmael_NNP ._.

Some_DT years_NNS ago_RB --_: never_RB mind_VB how_WRB long_JJ precisely_RB --_: having_VBG little_JJ or_CC no_DT money_NN in_IN my_PRP$ purse_NN ,_, and_CC nothing_NN particular_JJ to_TO interest_VB me_PRP on_IN shore_NN ,_, I_PRP thought_VBD I_PRP would_MD sail_VB about_IN a_DT little_JJ and_CC see_VB the_DT watery_JJ part_NN of_IN the_DT world_NN ._.

Results for noisy input (light photocopy):

'_POS cH___NNS '_POS R_`` l_NNS '_'' ._.

_omings_NNS ._.

,_, call_VBP me_PRP IshMael_NNP ._.

soMe_JJ ye_s_NNS ago_RB --_: never_RB mind_VB how_WRB long_JJ ,_, p_NNP ,_, ec_NNP ;_: sely__hav_NNP ;_: ng_NNP __NNP ;_: €tle_NNP or_CC no_DT _oney_NN in_IN my_PRP$ purse_NN ,_, and_CC nothing_NN p__NN ;_: ,_, u___JJ to_TO ;_: ,_, terest_NN Me_NN on_IN shore_NN ,_, I_PRP thoug_t_VBP I_PRP would_MD sail_VB _boUt_VBN a_DT __NN ;_: tt1e_JJ and_CC see_VBP _e_JJ watery_NN p__, or_CC the_DT world_NN ._.

Correspondence via Alignment

Idea: determine correspondences at each level of text processing by applying multiple levels of approximate string matching.

Well-known recurrence for string edit distance:

$$dist1_{0,0} = 0$$
$$dist1_{i,0} = dist1_{i-1,0} + c1_{del}(s_i)$$
$$dist1_{0,j} = dist1_{0,j-1} + c1_{ins}(t_j)$$

$$dist1_{i,j} = \min \begin{cases} dist1_{i-1,j} + c1_{del}(s_i) \\ dist1_{i,j-1} + c1_{ins}(t_j) \\ dist1_{i-1,j-1} + c1_{sub}(s_i, t_j) \end{cases}$$

By keeping track of optimal decision(s) at each step, we can trace back and recover correspondence (alignment) between two strings.

"A general method applicable to the search for similarities in the amino-acid sequences of two proteins," S. B. Needleman and C. D. Wunsch, *Journal of Molecular Biology*, vol. 48, 1970, pp. 443-453.

"The string-to-string correction problem," R. A. Wagner and M. J. Fischer, *Journal of the Association for Computing Machinery*, vol. 21, 1974, pp. 168-173.

Correspondence via Alignment

The traditional model for string matching only allows for single-symbol deletions, insertions, and substitutions.

As we have seen, however, the errors we face often involve faulty segmentation decisions (splits and merges).

E.g., m → rn

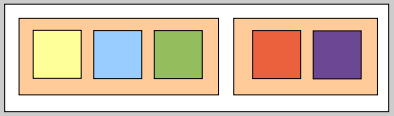
Update the recurrence to allow for generalized *k:l* substitutions:

$$dist1_{i,j} = \min \begin{cases} dist1_{i-1,j} + c1_{del}(s_i) \\ dist1_{i,j-1} + c1_{ins}(t_j) \\ \min_{1 \le k' \le k,\ 1 \le l' \le l} [\, dist1_{i-k',j-l'} + c1_{sub_{k:l}}(s_{i-k'+1 \ldots i}, t_{j-l'+1 \ldots j})] \end{cases}$$

Hierarchical Edit Distance

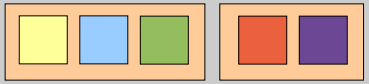
Traditional model will allow us to align any two sequences. To capture hierarchy, we apply three successive levels of matching:

Determine optimal correspondence between sentences by ...



... basic unit is sentences, made up of tokens.

... determining optimal corresondence between tokens by ...



... basic unit is tokens, made up of symbols ...

... comparing tokens allowing for deletions, insertions, substitutions, splits, and merges.

Basic unit is symbols...

When final correspondence determined, compare PoS tags as well.

Token Level Comparison

Looks similar to lowest-level comparison:

$$dist2_{i,j} = \min \begin{cases} dist2_{i-1,j} + c2_{del}(s_i) \\ dist2_{i,j-1} + c2_{ins}(t_j) \\ \min_{1 \le k' \le k,\ 1 \le l' \le l} [\ dist2_{i-k',j-l'} + \\ \qquad c2_{sub_{k:l}}(s_{i-k'+1\ldots i}, t_{j-l'+1\ldots j})] \end{cases}$$

Except now basic costs are defined in terms of that lower level:

$$c2_{del}(s_i) \equiv dist1(s_i, \phi)$$
$$c2_{ins}(t_j) \equiv dist1(\phi, t_j)$$
$$c2_{sub_{k:l}}(s_{i-k'+1\ldots i}, t_{j-l'+1\ldots j}) \equiv dist1(s_{i-k'+1\ldots i}, t_{j-l'+1\ldots j})$$

I.e., we are deleting, inserting, substituting, splitting, and merging tokens, not symbols.

Sentence Level Comparison

Looks similar to other two levels:

$$dist3_{i,j} = \min \begin{cases} dist3_{i-1,j} + c3_{del}(s_i) \\ dist3_{i,j-1} + c3_{ins}(t_j) \\ \min_{1 \le k' \le k,\ 1 \le l' \le l} [\ dist3_{i-k',j-l'} + \\ \qquad c3_{sub_{k:l}}(s_{i-k'+1\ldots i}, t_{j-l'+1\ldots j})] \end{cases}$$

Except now basic costs are defined in terms of second level:

$$c3_{del}(s_i) \equiv dist2(s_i, \phi)$$
$$c3_{ins}(t_j) \equiv dist2(\phi, t_j)$$
$$c3_{sub_{k:l}}(s_{i-k'+1\ldots i}, t_{j-l'+1\ldots j}) \equiv dist2(s_{i-k'+1\ldots i}, t_{j-l'+1\ldots j})$$

I.e., we are deleting, inserting, substituting, splitting, and merging sentences, not tokens or symbols.

Test Conditions

Corpus	10 pages of Project Gutenberg *Moby-Dick*. http://www.gutenberg.net (Michael Hart et al.)
Optical character recognition	Open Source *gocr* package. http://jocr.sourceforge.net/index.html (Joerg Schulenburg et al.)
Sentence boundary detection	MXTERMINATOR. "A Maximum Entropy Approach to Identifying Sentence Boundaries," J. C. Reynar and A. Ratnaparkhi, *Proc. 5th Conf. on Applied Natural Language Processing*, 1997.
Tokenization	Penn Treebank tokenizer. http://www.cis.upenn.edu/~treebank/tokenizer.sed (Robert MacIntyre)
Part-of-speech tagging	MXPOST. "A Maximum Entropy Part-Of-Speech Tagger," A. Ratnaparkhi, *Proc. Empirical Methods in Natural Language Processing Conference*, 1996.

Test Conditions (cont.)

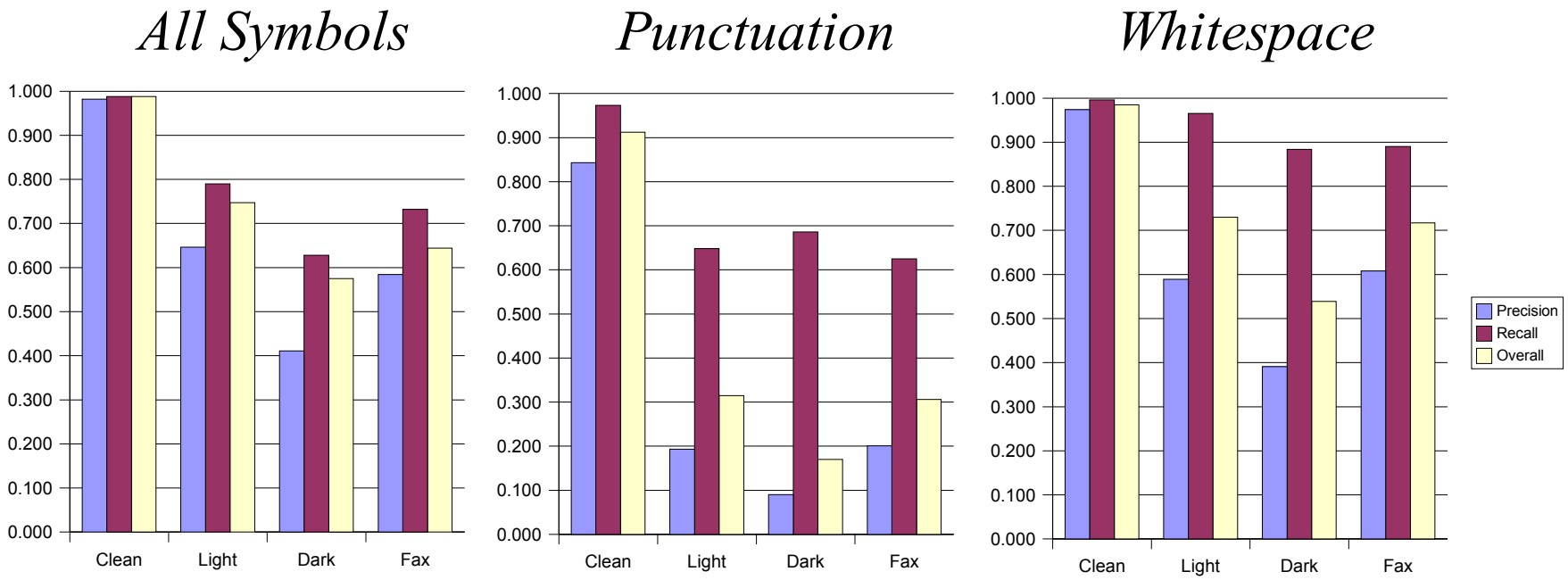
Corpus text was:

- formatted in 12-point Times font using MS Word;
- printed using laserprinter;
- used to create four test sets: one used as-is ("clean"), one copied light ("light"), one copied dark ("dark"), one faxed ("fax");
- scanned at 300 dpi.

Important note: current study is not an attempt to evaluate the text processing algorithms. We are evaluating the evaluation paradigm:

- Does it provide useful measures of accuracy?
- Can it recover correct correspondences for use in later analyses?

Average OCR Performance



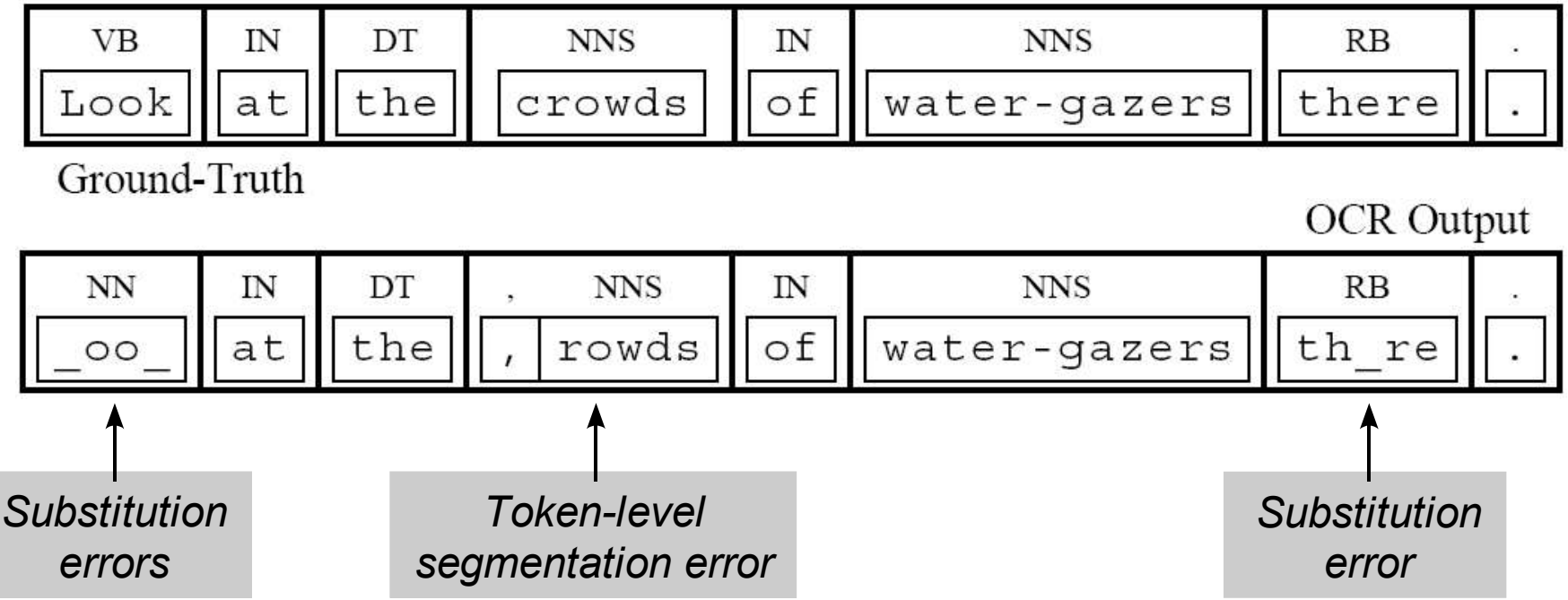
Notes:

- Baseline high on clean inputs, deteriorates rapidly on noisy inputs.
- Punctuation especially badly impacted: many false alarms.

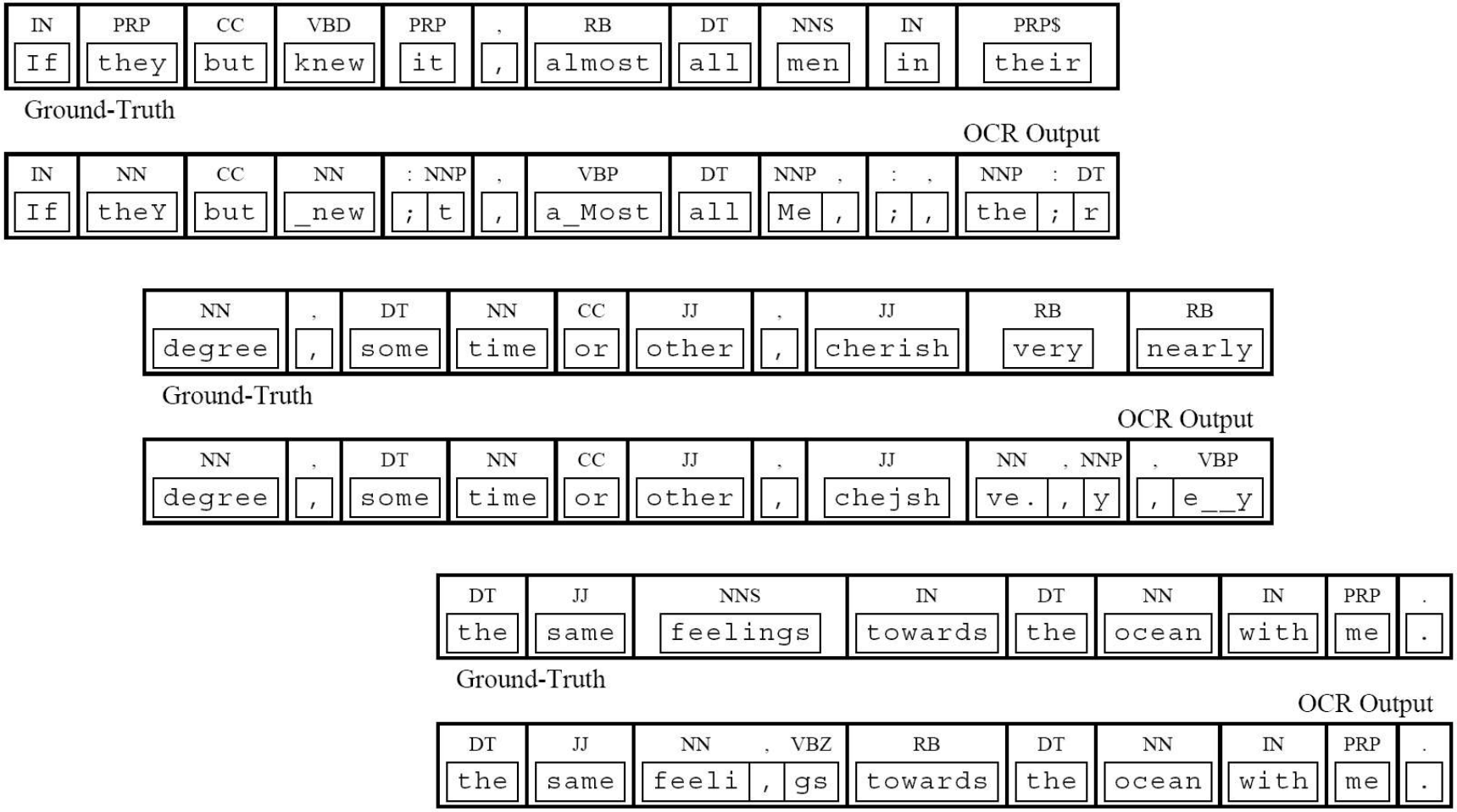
Sample Alignment 1

Applying hierarchical string matching paradigm, we can recover correct correspondence between noisy output and original input.

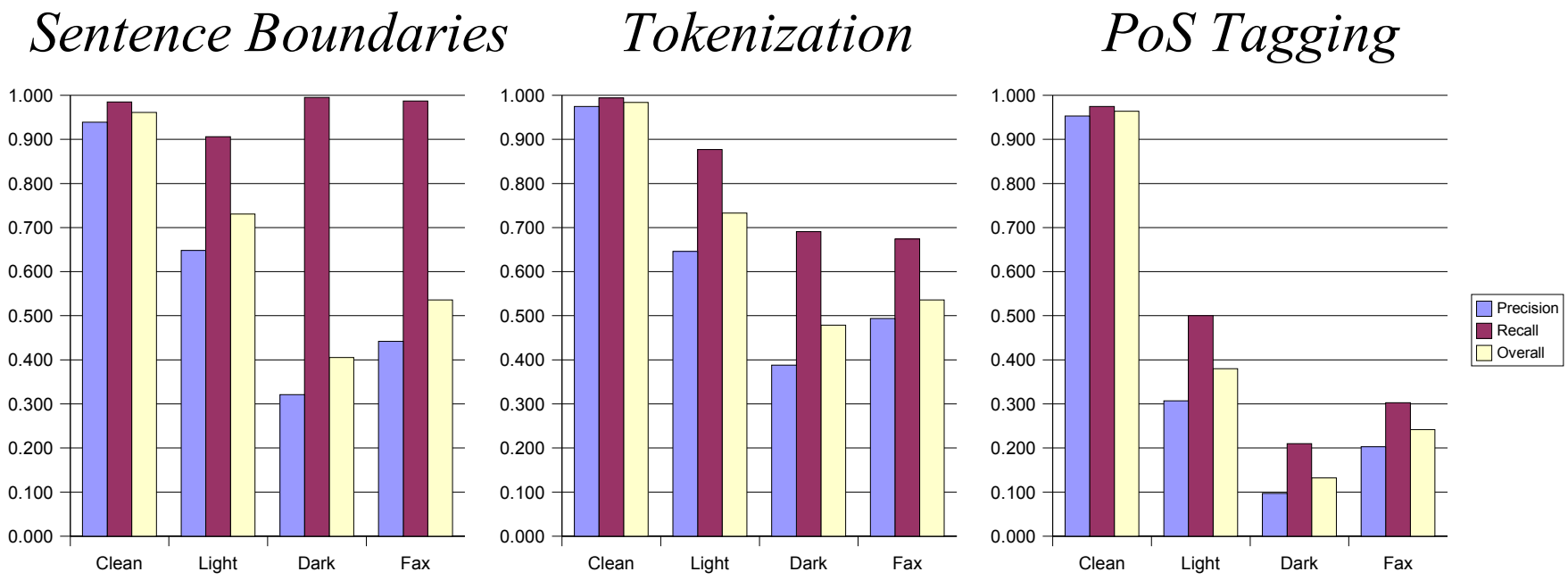
A straightforward example found by algorithm:



Sample Alignment 2



Text Processing Performance



Notes:

- Clean input processed at > 95%; many false alarms in noisy inputs.
- Performance degrades with each successive stage.

Conclusions

Proposed approach for analyzing impact of OCR errors on text processing seems effective:

- Provides formalism for identifying and visualizing errors.
- Allows performance to be quantified in fine-grained way.

Future work:

- Identify specific classes of errors that have largest impact.
- Address through more accurate document analysis and OCR.
- Study whether text processing can also be made more robust.
- For final end-user applications, develop interface and interaction paradigms to help user cope with imperfect data.