ParaMetric:

An Automatic Evaluation Metric for Paraphrasing

Chris Callison-Burch
Trevor Cohn Mirella Lapata

Motivation for Paraphrasing

- Useful in NLP application such as:
  - **Generation** - creating more varied and fluent text
  - **Question answering** - identifying passages that match the expected answer type
  - **Multi-document summarization** - condensing repeated information
- Many recent data-driven approaches
  - Distributional similarity of dependencies (Lin and Pantel)
  - Aligning multiple translations of a foreign text (Barzilay and McKeown, Pang et al)
  - Statistical translation applied to comparable corpora (Quirk et al)
  - Pivoting through foreign phrases (Bannard and Callison-Burch)

Previous approaches to evaluation

- Subjective evaluation
  - Ask judges whether paraphrases
    - “are roughly interchangeable given the genre”
    - “are approximately conceptually equivalent”
    - “preserve meaning and remain grammatical”
- Task-based
  - Apply paraphrases to another task with its own evaluation metric
    - Machine translation
    - Question answering
    - Query expansion

Goals of ParaMetric

- Automatic evaluation of paraphrases
  - Objective
  - Repeatable
  - Low cost
  - Task-independent
- Only aiming to evaluate lexical and phrasal paraphrases, not sentential ones

Challenges for Automatic Evaluation

- Developing exhaustive list of paraphrases is difficult / impossible
- Calculating precision and recall is problematic if the list of valid paraphrases is incomplete

$$Prec = \frac{|G \cap A|}{|A|}$$ ← underestimate

$$Recall = \frac{|G \cap A|}{|G|}$$ ← relative to incomplete set

Challenges for Automatic Evaluation

- Context determines whether a particular paraphrase is valid

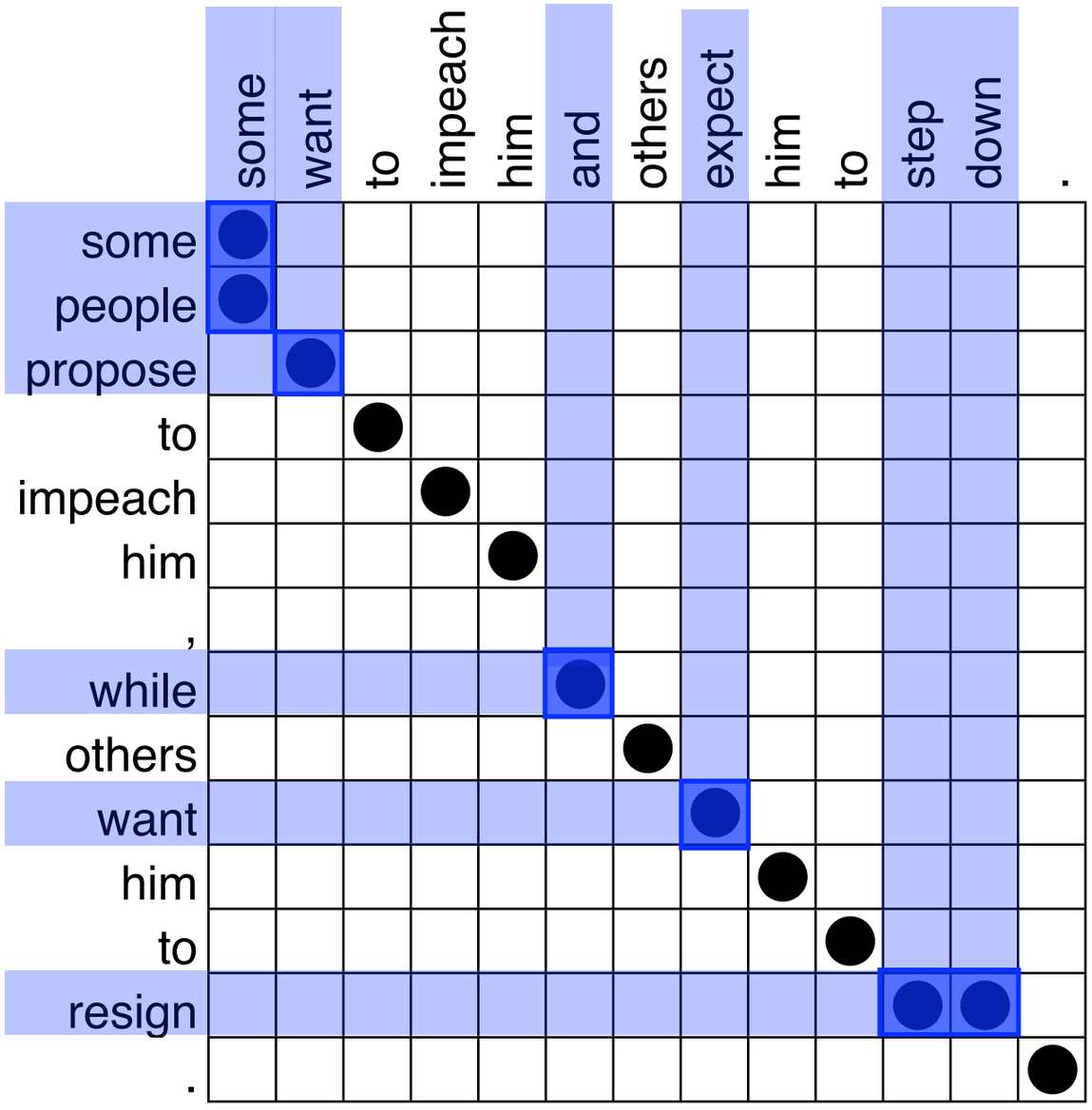
Emma cried and he tried to **comfort** her.
Emma cried and he tried to **console** her.

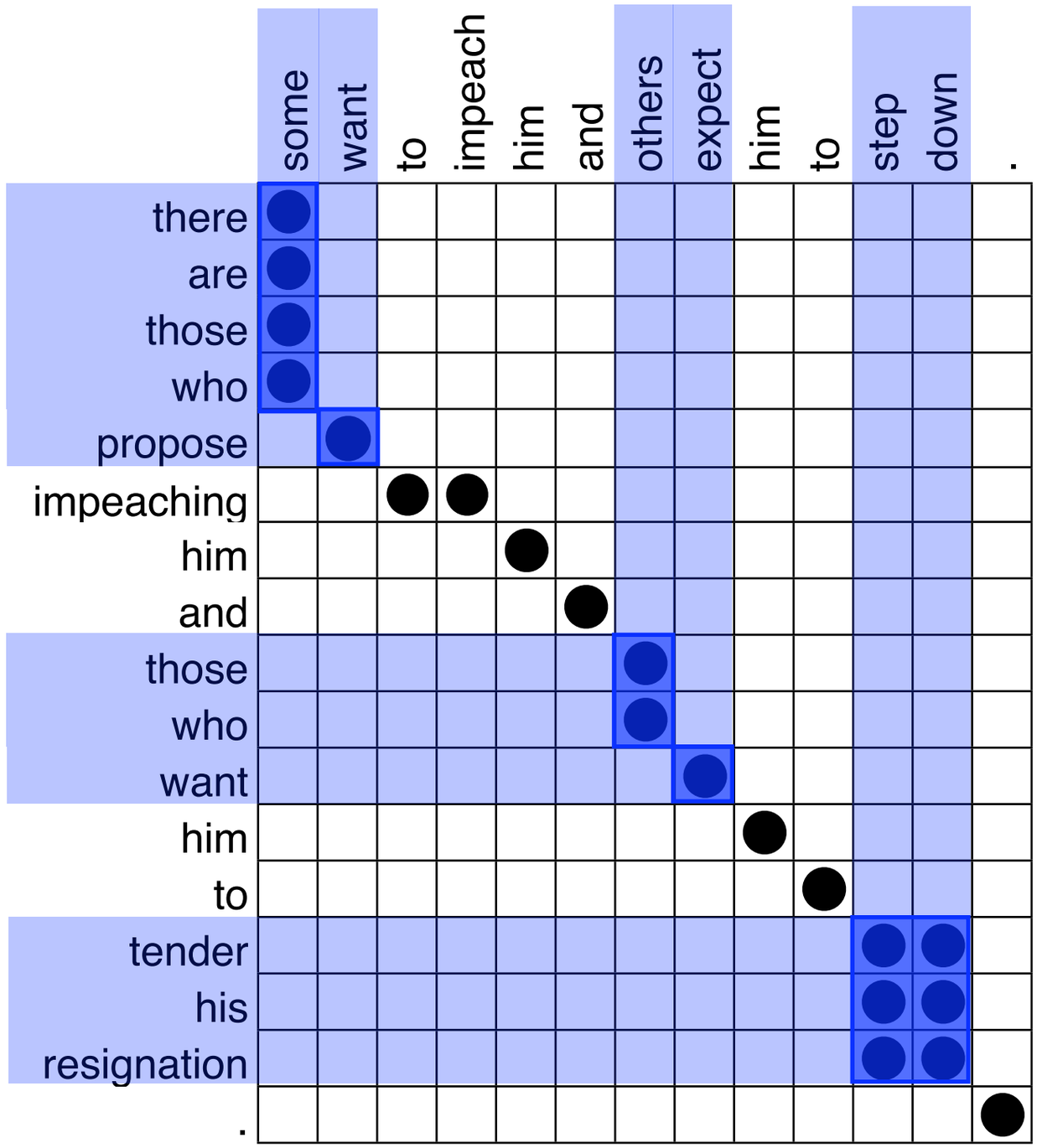
George Bush said Democrats provide **comfort** to our enemies.
George Bush said Democrats provide **console** to our enemies.

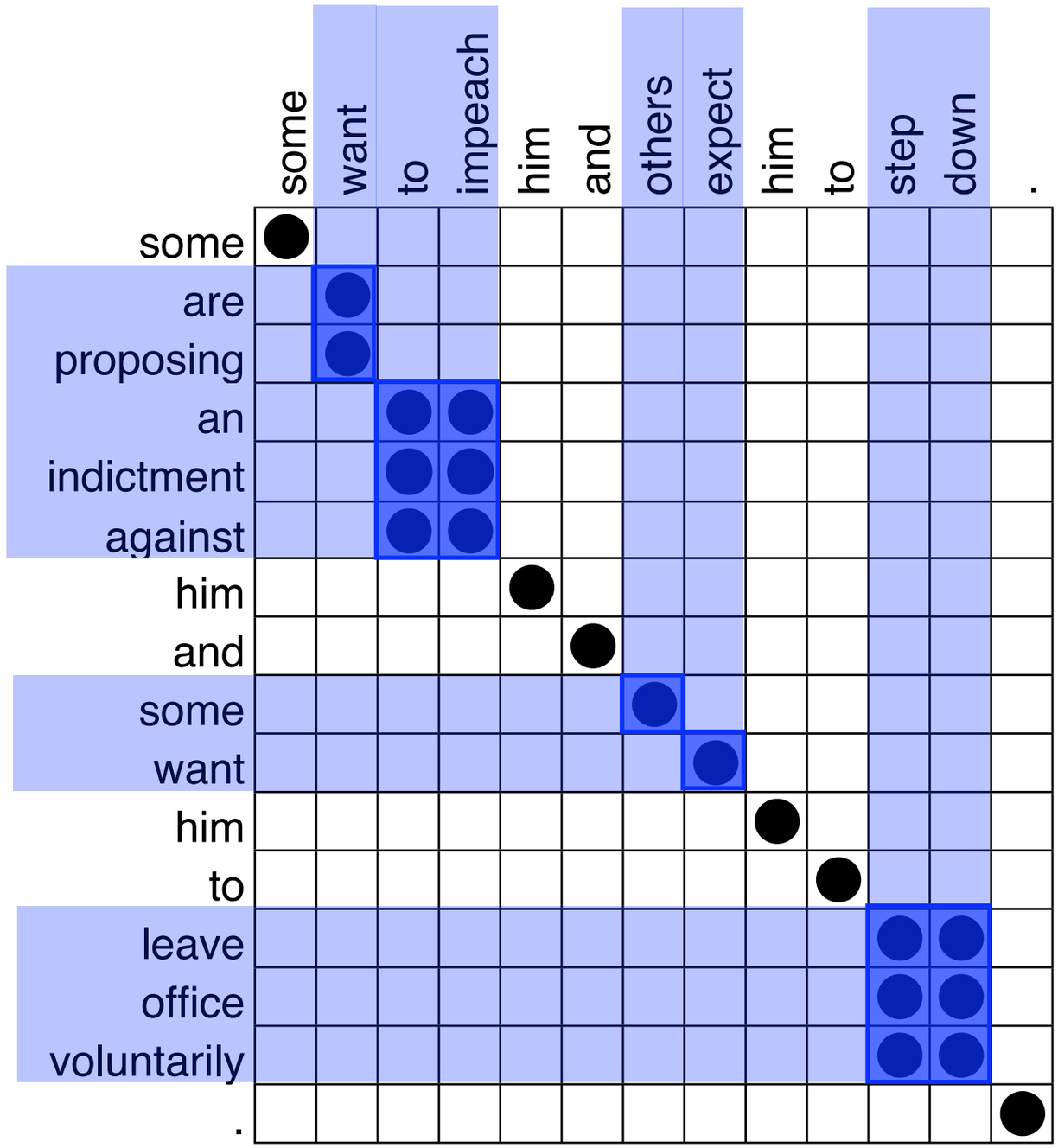
Aligning Paraphrases

- Develop gold standard paraphrases
  - Manually align correspondences in equivalent sentences
  - Drawn from multiple reference translations
  - Extract paraphrases from manual alignments

Some want to impeach him and others expect him to step down.
Some people propose to impeach him, while others want him to resign.
There are those who propose impeaching him and those who want him to tender his resignation.
Some are proposing an indictment against him and some want him to leave office voluntarily.

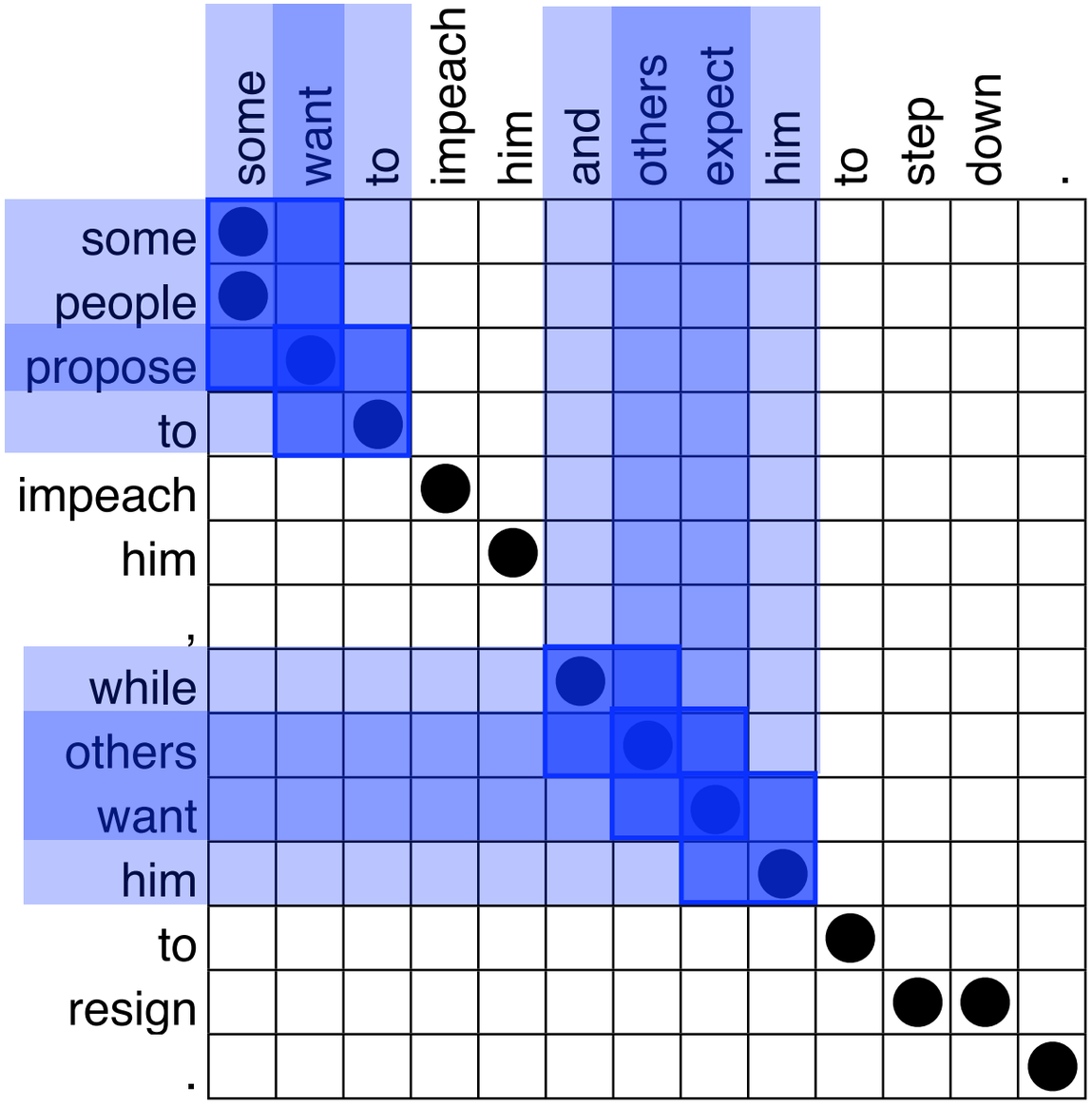

some
want
to
impeach
him
and
others
expect
him
to
step
down
.
some
people
propose
to
impeach
him
,
while
others
want
him
to
resign
.

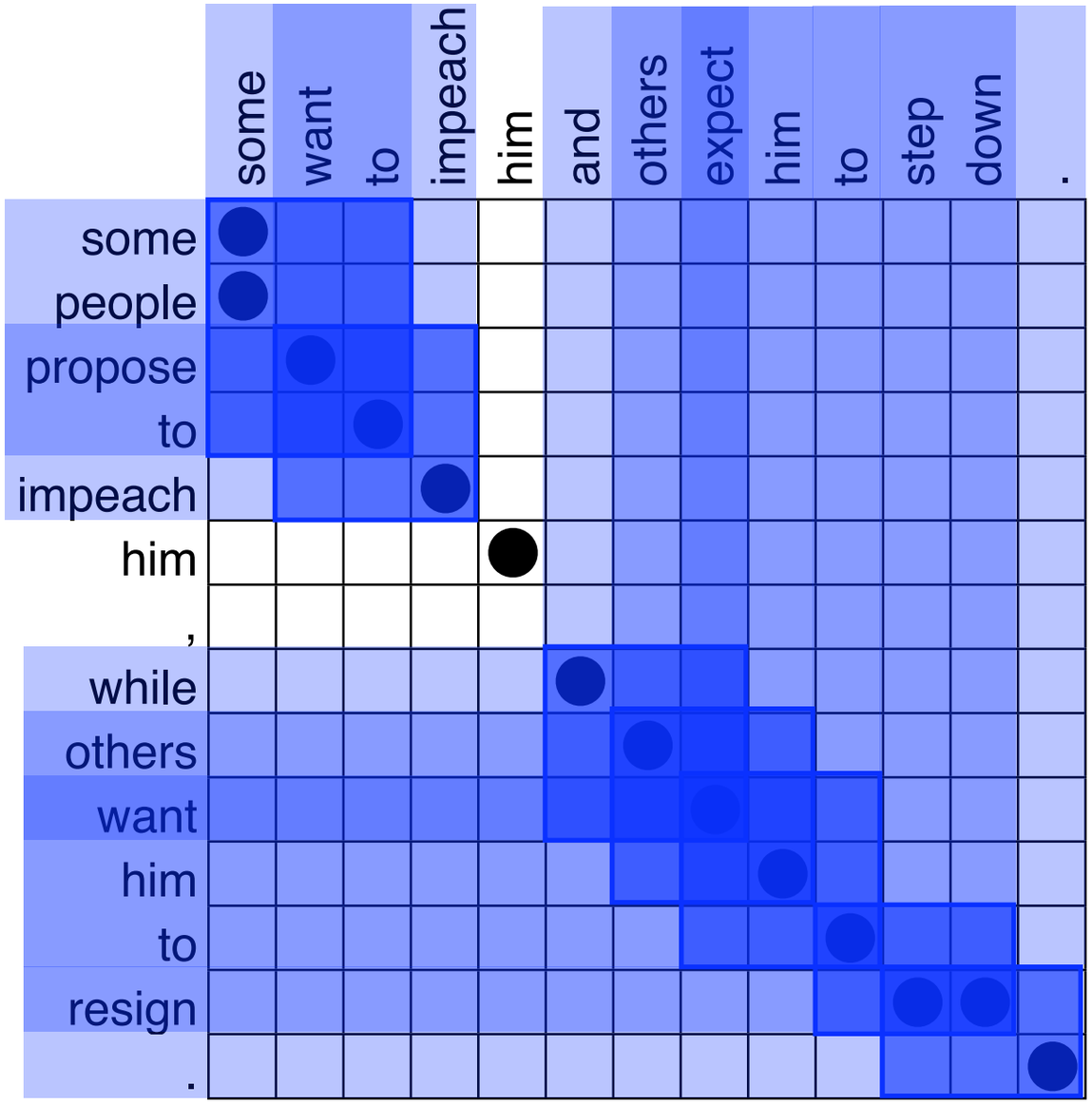

some
want
to
impeach
him
and
others
expect
him
to
step
down
.
there
are
those
who
propose
impeaching
him
and
those
who
want
him
to
tender
his
resignation
.

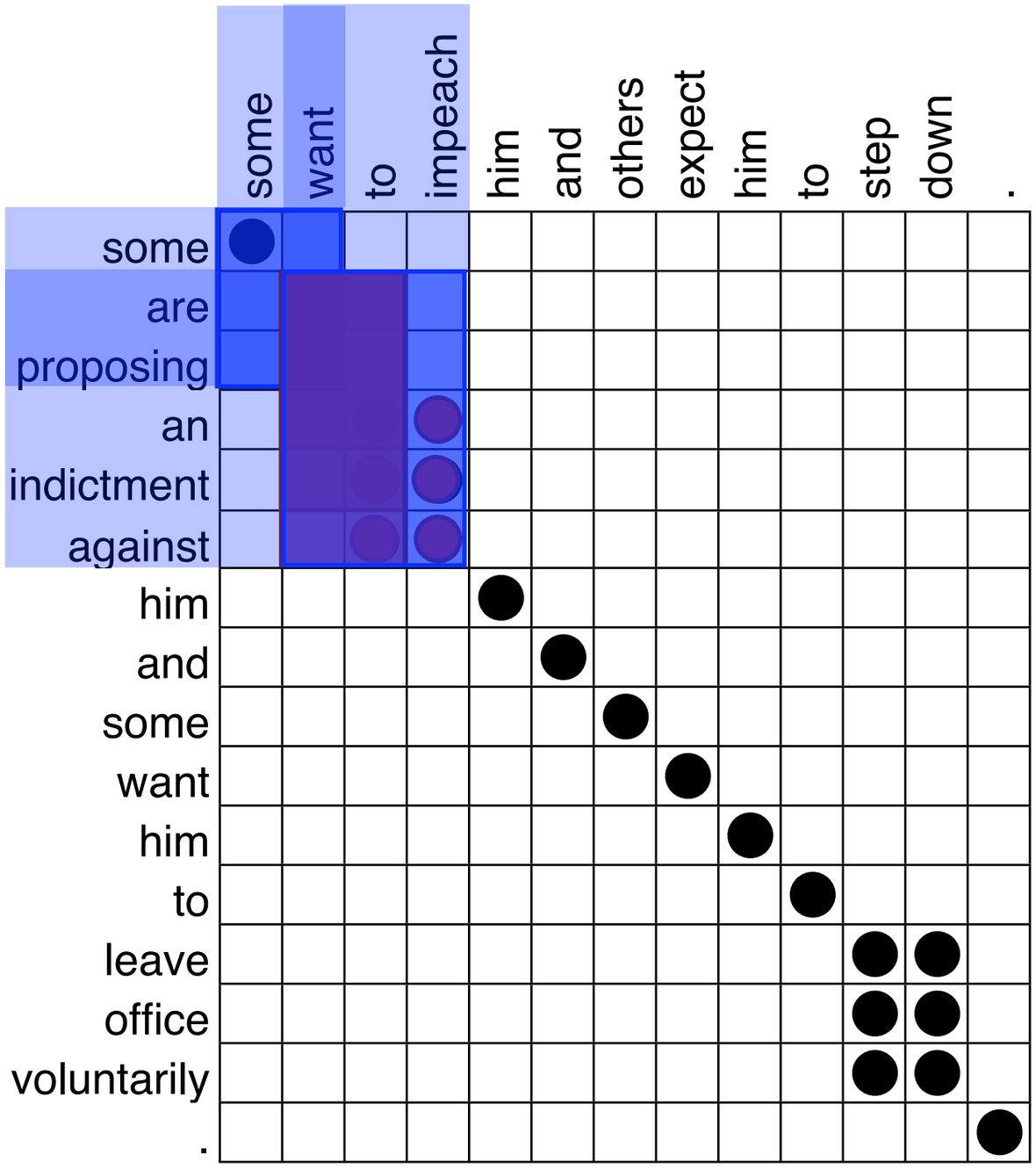

some
want
to
impeach
him
and
others
expect
him
to
step
down
.
some
are
proposing
an
indictment
against
him
and
some
want
him
to
leave
office
voluntarily
.

Basic paraphrases

some	some people, there are those who
want	propose, are proposing
to impeach	an indictment against, impeaching
and	while
others	some, those who
expect	want
step down	resign, leave office voluntarily, tender his resignation


some
want
to
impeach
him
and
others
expect
him
to
step
down
.
some
people
propose
to
impeach
him
,
while
others
want
him
to
resign
.


some
want
to
impeach
him
and
others
expect
him
to
step
down
.
some
people
propose
to
impeach
him
,
while
others
want
him
to
resign
.


some
want
to
impeach
him
and
others
expect
him
to
step
down
.
some
are
proposing
an
indictment
against
him
and
some
want
him
to
leave
office
voluntarily
.

Extended paraphrases

some want	some people propose, there are those who propose, some are proposing
some want to	some people propose to
some want to impeach	some people propose to impeach, there are those who propose impeaching, some are proposing an indictment against
want to	propose to
want to impeach	propose to impeach, propose impeaching, are proposing an indictment against
...	...

- A total of 142 non-identical phrase pairs can be extracted from the 3 sentence pairs

Annotated Data

- Manually aligned multiple reference translations from Chinese-English set
  - Chinese translated by 11 different human translators
  - The first reference was paired with each of the other 10 (i.e. 1-2, 1-3, 1-4, ..., 1-11)
  - A total of 500 pairs were annotated
  - Started with automatic alignments and corrected them
  - Avg annotation time was 77 seconds for a total of 11 hours worth of effort
- 14,000+ unique paraphrases containing 5 words or less

Evaluation Metrics

- Two types of metrics

Evaluation Metrics

- Two types of metrics
- First type:
  - ***Alignment precision*** and ***alignment recall***
  - Paraphrasing method creates alignments between sentence pairs similar to the manual alignments
  - Precision and recall can be accurately calculated

Evaluation Metrics

- Two types of metrics
- First type:
  - ***Alignment precision*** and ***alignment recall***
  - Paraphrasing method creates alignments between sentence pairs similar to the manual alignments
  - Precision and recall can be accurately calculated
  - Restricted to methods that can align sentences

Evaluation Metrics

- Two types of metrics
- First type:
  - ***Alignment precision*** and ***alignment recall***
  - Paraphrasing method creates alignments between sentence pairs similar to the manual alignments
  - Precision and recall can be accurately calculated
  - Restricted to methods that can align sentences
- Second type:
  - Enumerate paraphrase lists from manual alignments
  - Calculate precision and recall against these lists
  - Can be applied to any paraphrasing method

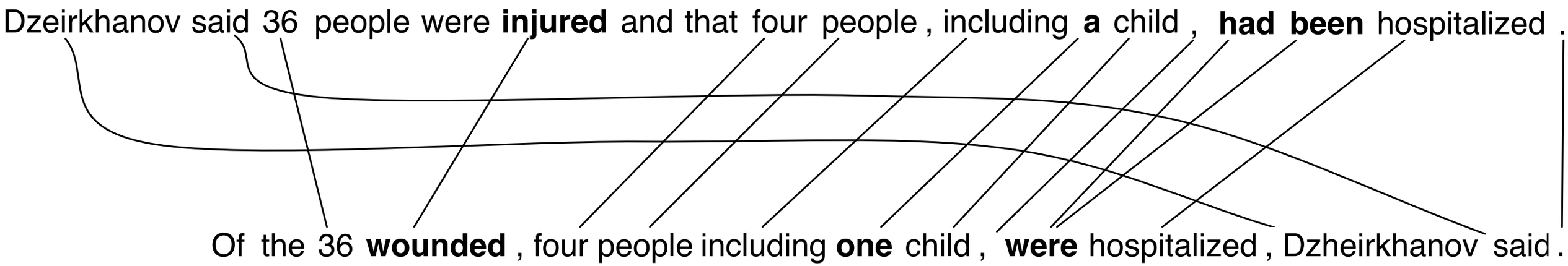
Evaluation Metrics

- Two types of metrics
- First type:
  - ***Alignment precision*** and ***alignment recall***
  - Paraphrasing method creates alignments between sentence pairs similar to the manual alignments
  - Precision and recall can be accurately calculated
  - Restricted to methods that can align sentences
- Second type:
  - Enumerate paraphrase lists from manual alignments
  - Calculate precision and recall against these lists
  - Can be applied to any paraphrasing method
  - Provides ***lower bound on precision*** and a ***recall estimate relative to test set***

Applied to three different paraphrasing methods

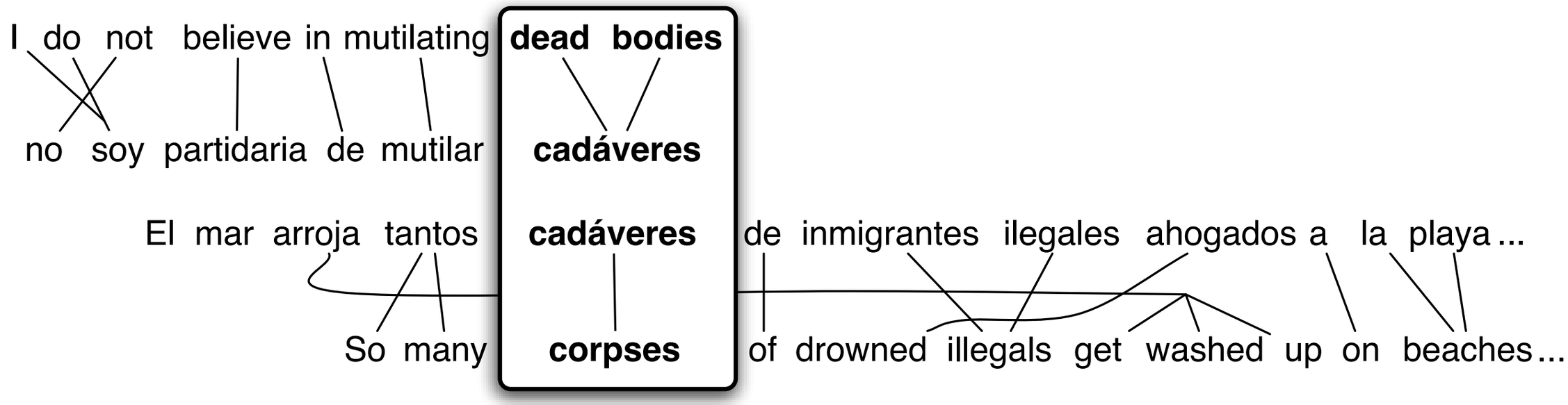
- Monolingual SMT (Quirk et al 2004)
- Pivoting through foreign phrases in a bilingual corpus (Bannard and Callison-Burch 2005)
- Syntactic alignment (Pang et al 2003)

Monolingual SMT



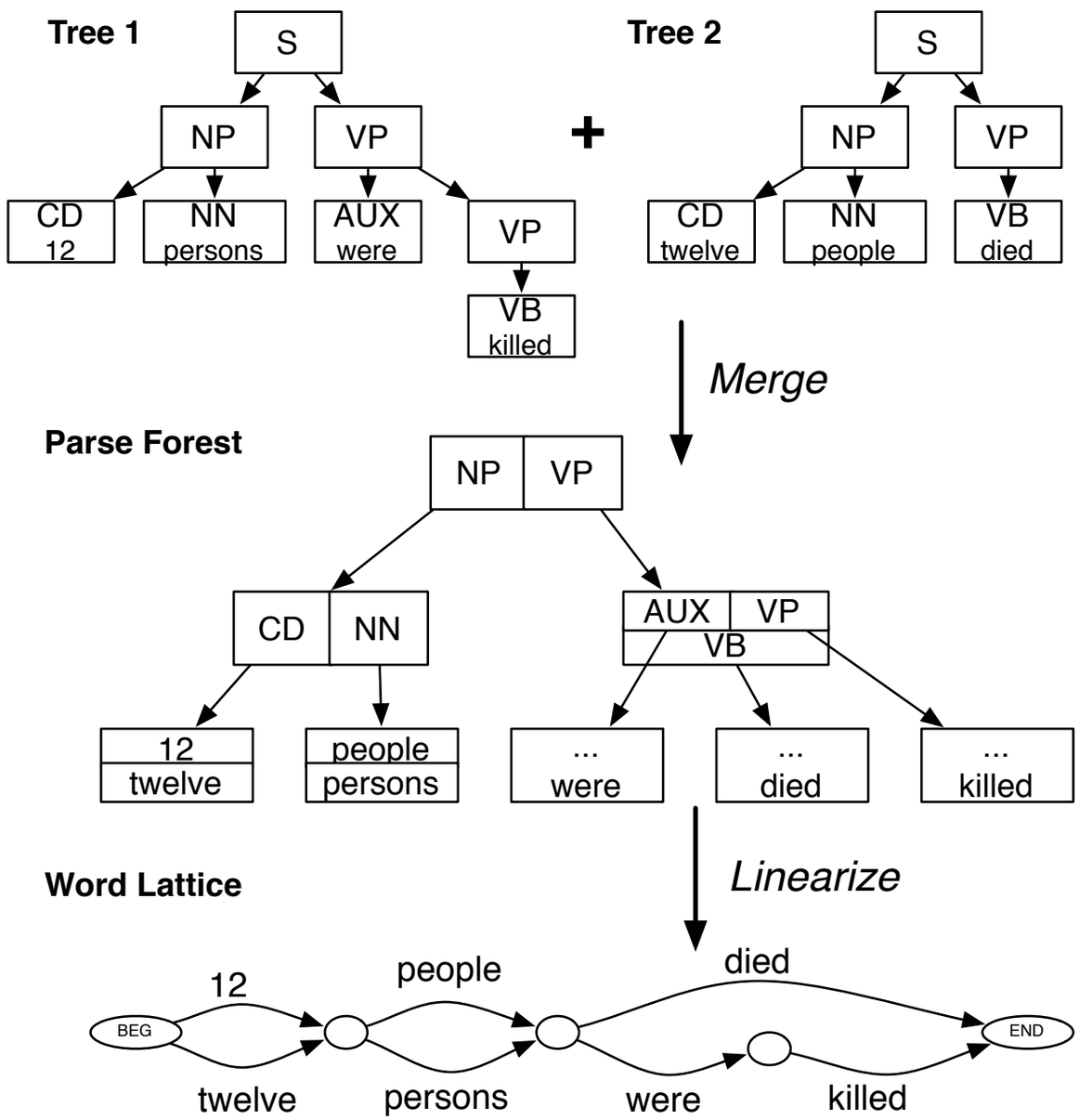
- Trained on 100k sentence pairs / 3 million words worth of paired reference translations for AlignPrec and AlignRecall
- Used MSR Paraphrase Phrase Table for LB-Prec and Rel-Recall

Bilingual Corpora



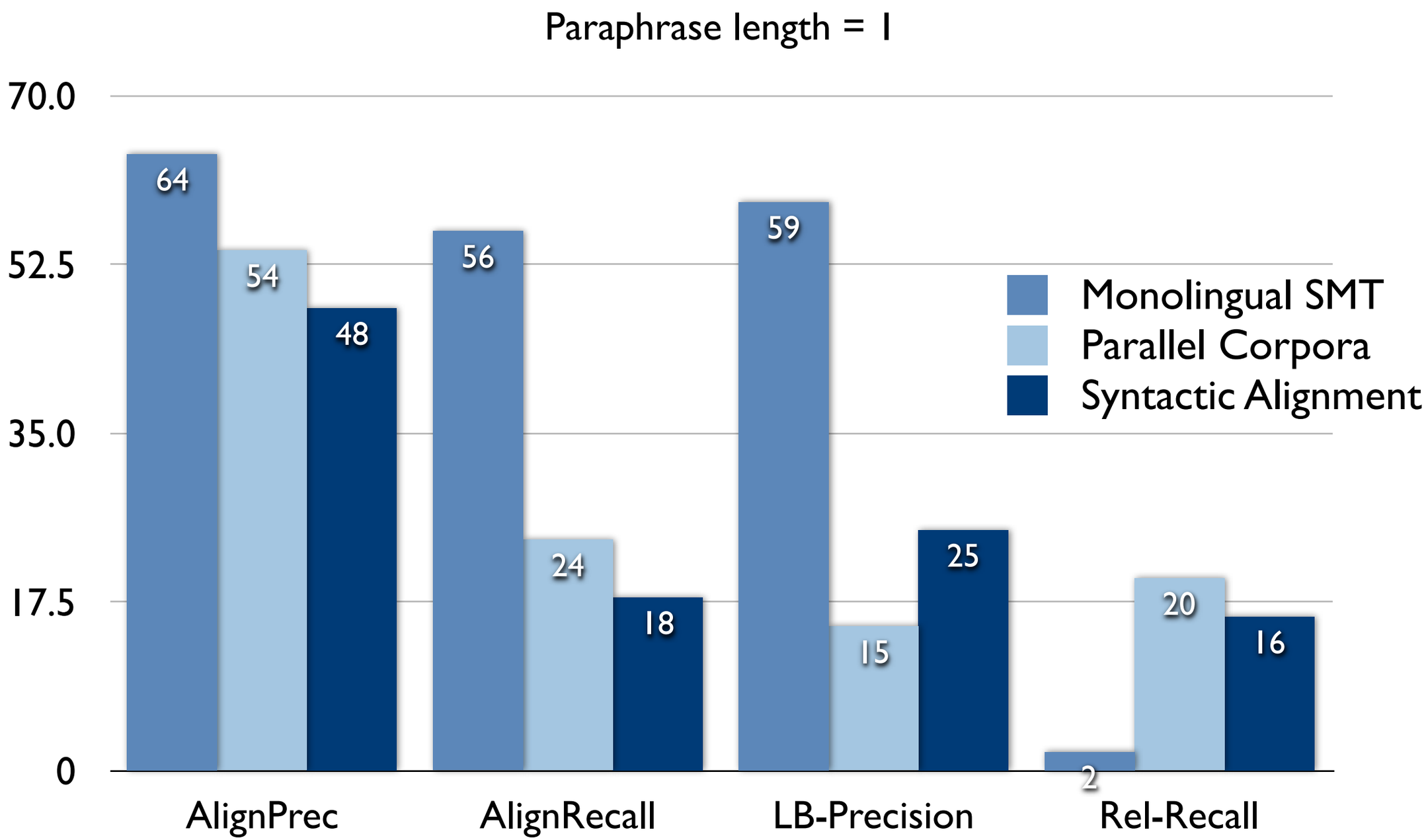
- Trained on 40 million word Chinese-English parallel corpus for AlignPrec and AlignRecall
- Trained on 10 Europarl parallel corpora for LB-Prec and Rel-Recall

Syntactic Alignment

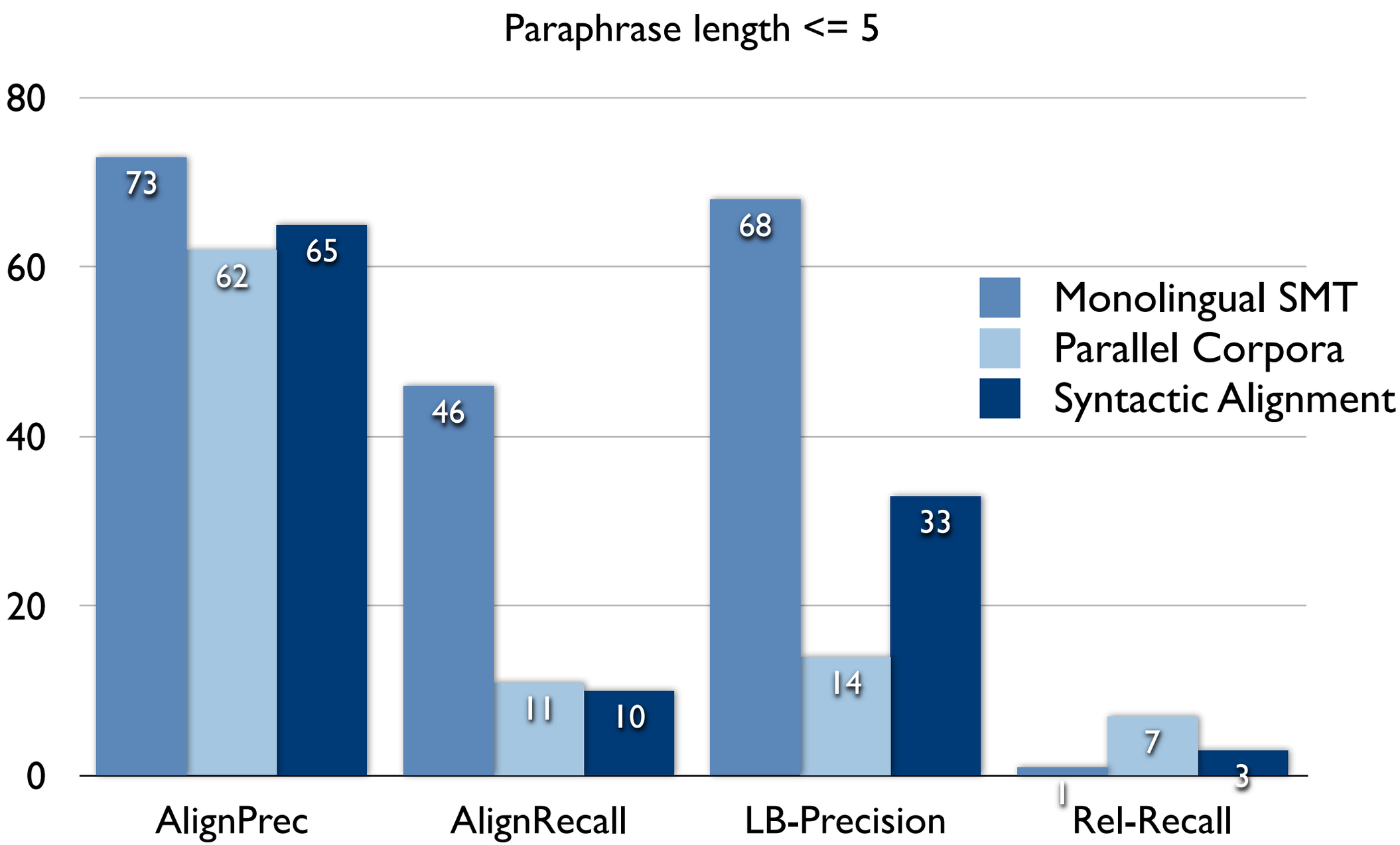


- Trained on full set of multiple translations for AlignPrec & AlignRecall
- Test set excluded for LB-Prec and Rel-Recall

Results


Paraphrase length = 1
70.0
52.5
35.0
17.5
0
64
54
48
56
24
18
59
15
25
2
20
16
AlignPrec
AlignRecall
LB-Precision
Rel-Recall
Monolingual SMT
Parallel Corpora
Syntactic Alignment

Results


Paraphrase length <= 5
80
60
40
20
0
73
62
65
46
11
10
68
14
33
1
7
3
AlignPrec
AlignRecall
LB-Precision
Rel-Recall
Monolingual SMT
Parallel Corpora
Syntactic Alignment

Conclusions

- Presented an automatic evaluation metric for paraphrasing
- Overcomes problem of incomplete list of paraphrases by redefining the problem as one of alignment
- Overcomes problem of context by using extracting paraphrases from equivalent sentences
- Allows repeatable, task-independent, objective measure of paraphrase quality

Shameless Plug

- We’ve got an article appearing in Computational Linguistics which details the creation of the data and its other uses.
- Coming soon:

  Constructing Corpora for the Development and Evaluation of Paraphrase Systems

Thanks!