Reassessment of the Role of Phrase Extraction in PBSMT

Francisco Guzmán
CCIR-ITESM
guzmanhe@gmail.com

Qin Gao
LTI-CMU
qing@cs.cmu.edu

Stephan Vogel
LTI-CMU
stephan.vogel@cs.cmu.edu

Presented by:
Nguyen Bach
LTI-CMU
nbach@cs.cmu.edu

Outline

- **Introduction**
- **Analysis Setup**
- **Analysis I: Word Alignments**
  - Metrics
  - Results
- **Analysis II: Phrase Extraction**
  - Metrics
  - Manual Evaluation
  - Results
- **Lessons Learned: Mind your gaps**
  - Experimental Results
- **Conclusions**

Word Alignment

- Beginning of SMT pipeline.
- Most subsequent steps based on WA.
- A lot of work to improve WA quality.
- Widely available Hand Alignments enabled discriminative approaches.
- Discriminative based on metrics such AER.

AER vs BLEU

- Fraser and Marcu, 2004

$$AER \neq BLEU$$

  - Evaluated **correlation** between BLEU and AER.
  - Possible links => flaws.
  - Variation of **F-measure**, uses a coefficient to **modify balance** between **precision** and **recall**.
  - The optimal **coefficient** depends on the **corpus**.
- Vilar et al., 2004
  - **Better BLEU** scores can be obtained with **"degraded" alignments.**
  - **Mismatch** between **alignment** and **translation** models.
  - Support the **use of AER**.

$$\downarrow AER \Rightarrow \uparrow BLEU$$

AER vs BLEU

- Ayan and Dorr, 2004
  - Analyze the quality of the alignments and resulting phrase tables.
  - Several types of alignments.
  - Several lexical weightings
  - CPER (Consistent Phrase Error Rate)

Our Objective

- Study the relationship between **Word Alignment** and **Phrase Extraction**
  - Alignment characteristics => Phrase Table characteristics

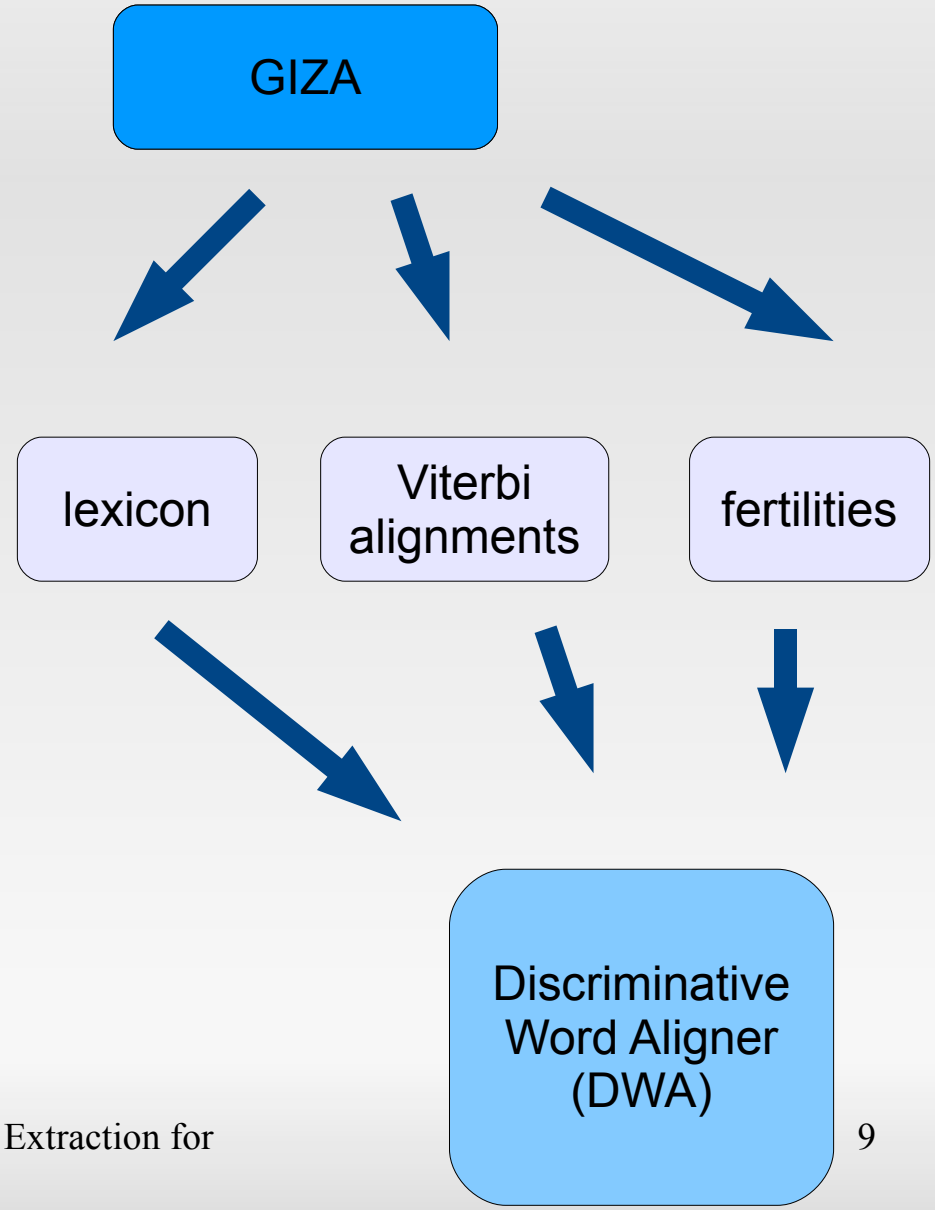
Analysis Setup

Setup

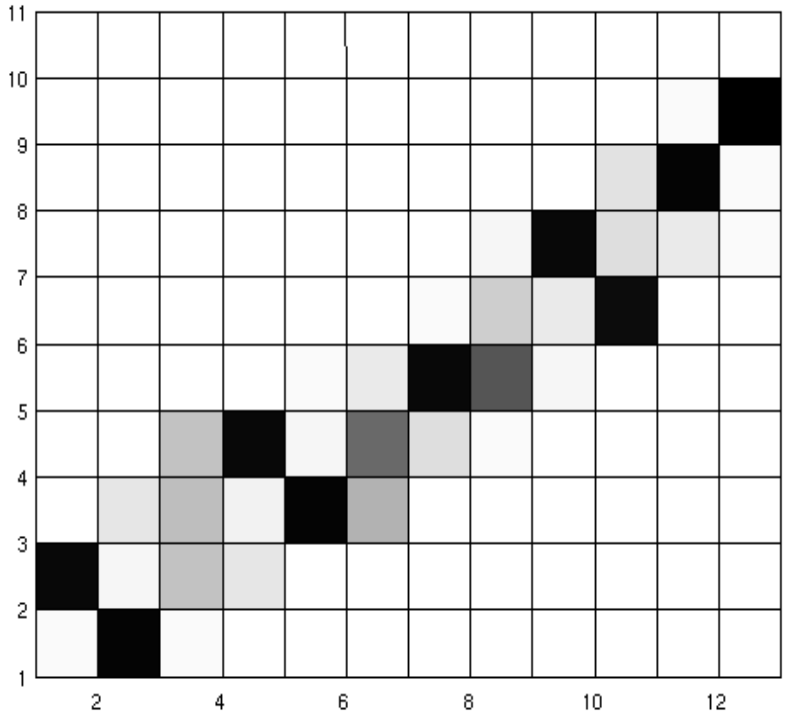
- Alignments:
  - GIZA++
  - Symmetrized (grow-diag-final)
  - Discriminative
- Phrase Extraction:
  - phrase-extract (Och & Ney)
  - Max P Len= 7
- Data:
  - Chinese-English
  - GIZA++ train (11M Sentences)
  - DWA tune(500 Sentences)
  - Test (200 Sentences)

Discriminative (Niehues & Vogel)

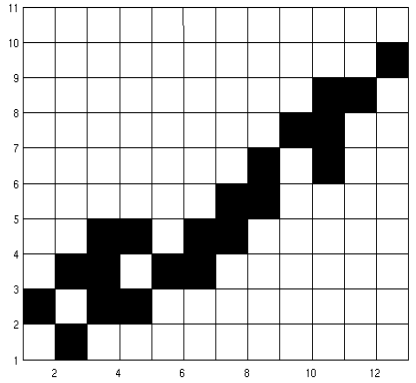
- CRF Based
- Uses GIZA++ byproducts as features
- Tuned towards AER



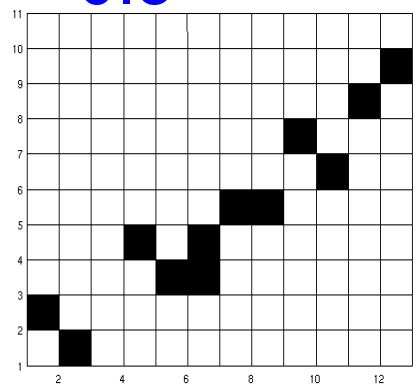
Continuous alignment matrix



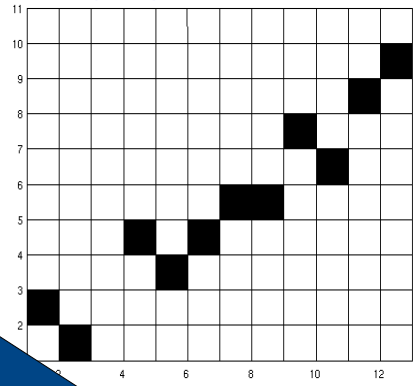
0.1



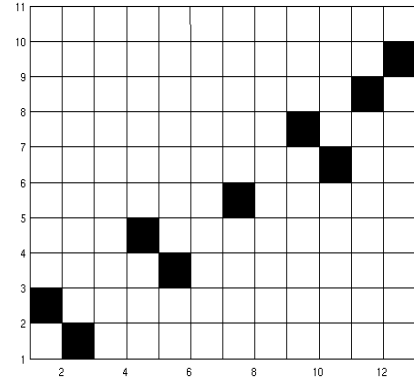
0.3



0.5



0.7



Binarized to different thresholds

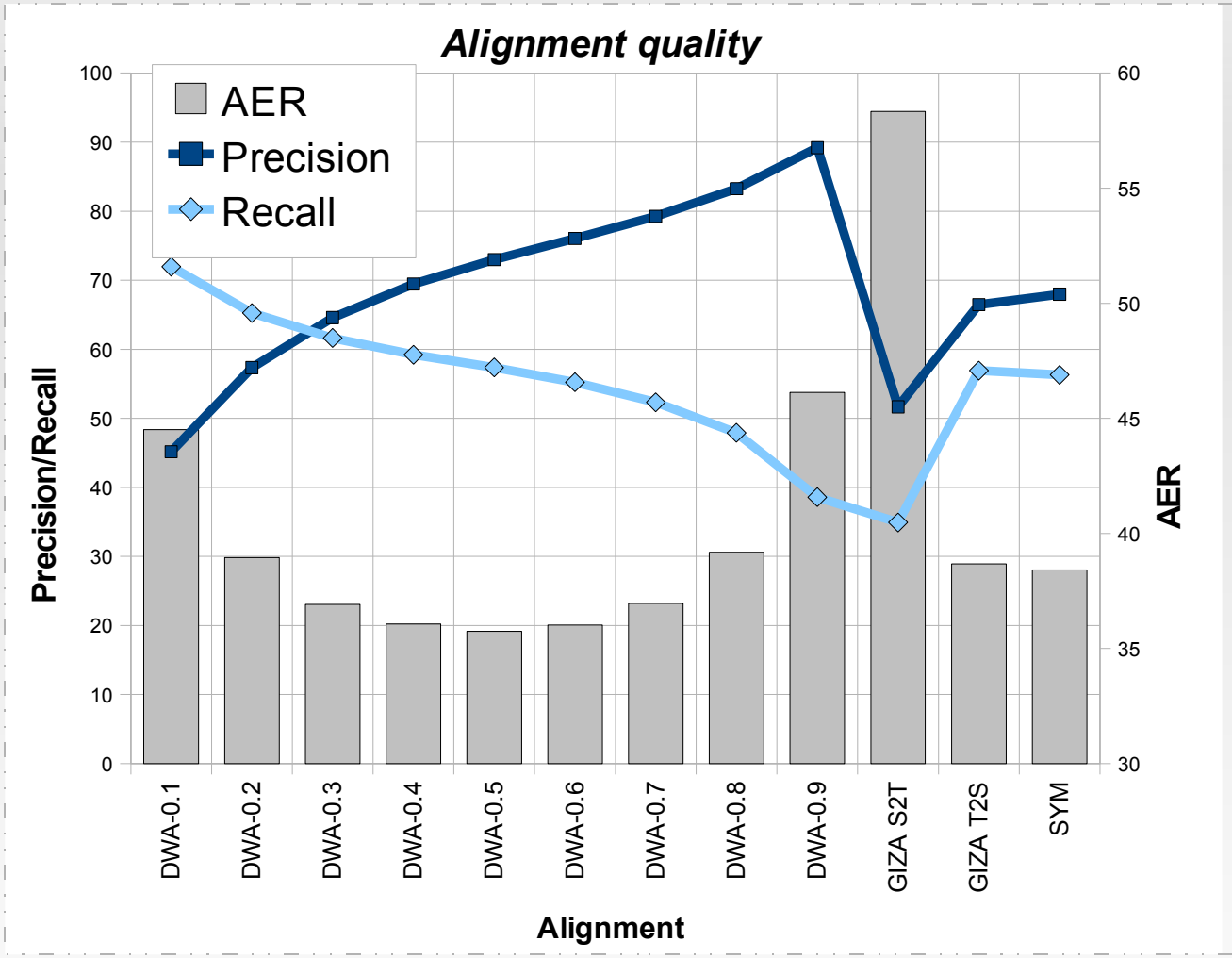
Analysis I: Word Alignment

Word Alignment Metrics

- Qualitative:
  - AER (F-measure)
  - Precision
  - Recall
- Quantitative:
  - Number of links
  - Number of unaligned words

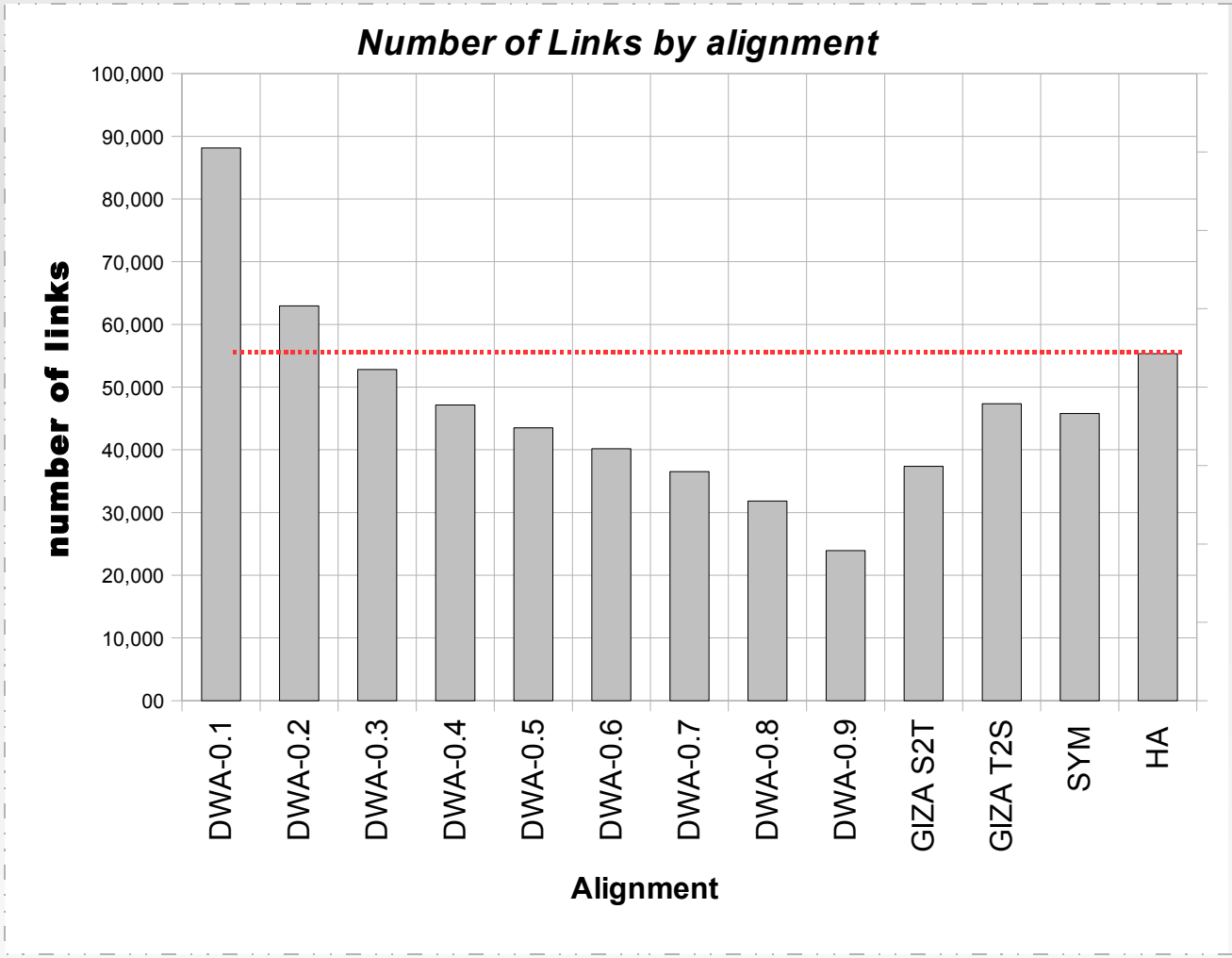
Alignment quality results

- DWA alignments: higher threshold=>more precision
- Best AER from slightly more precise alignment (DWA-0.5)
- GIZA=> more precision than recall.
- SYM lower AER than GIZA.



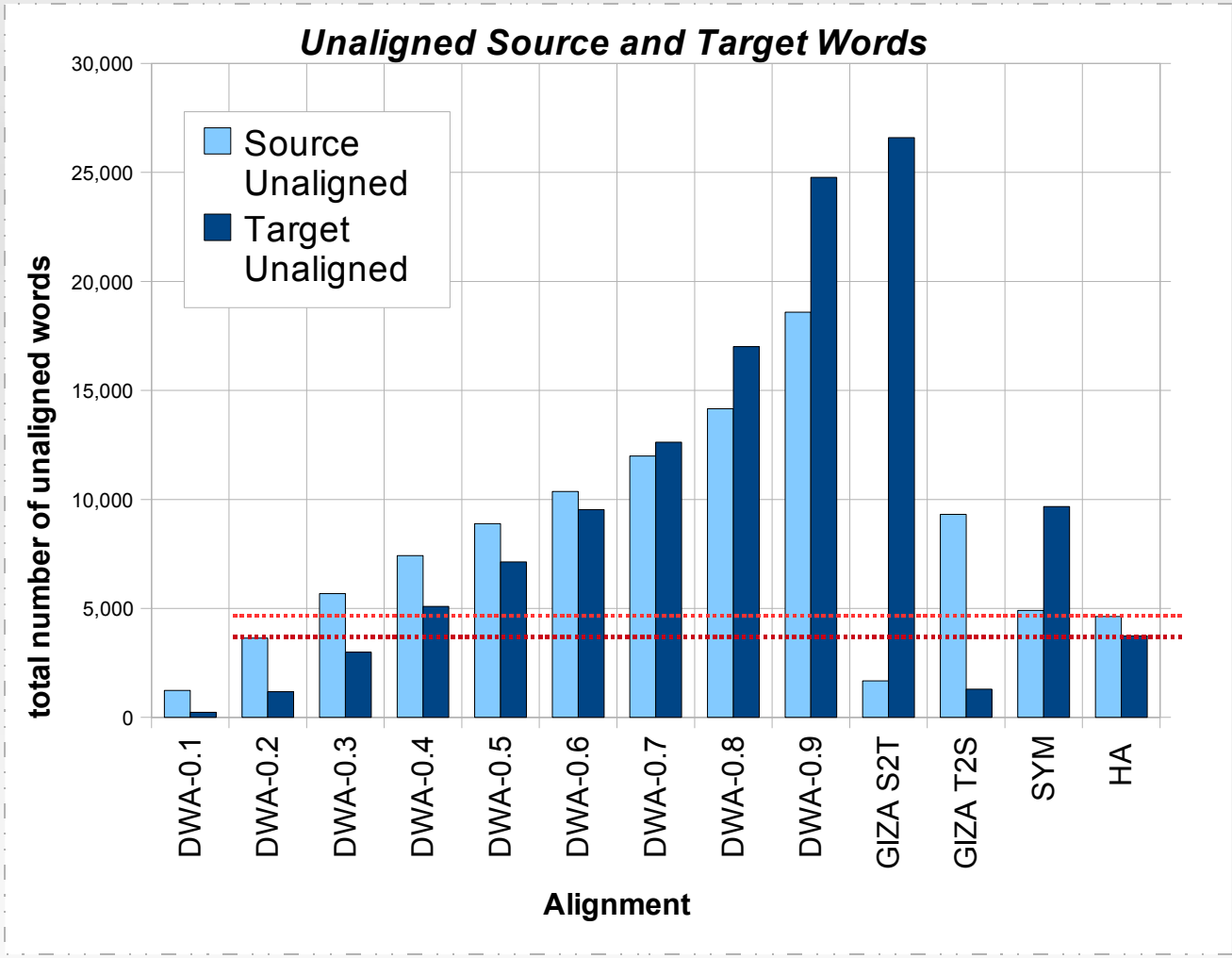
Links

- Hand Align closer to (DWA-0.3)
- DWA aligns: high threshold=>fewer links
- Best AER (DWA-0.5) fewer links than HA



Unaligned Words

- HA Source: closer to SYM, DWA-0.3
- HA Target: closer to DWA-0.4, DWA-0.3
- GIZA asymmetry
- DWA: higher threshold, more unalignments.
- DWA: lower threshold=> more proportion Chinese words unaligned.



Word Alignment: Summary

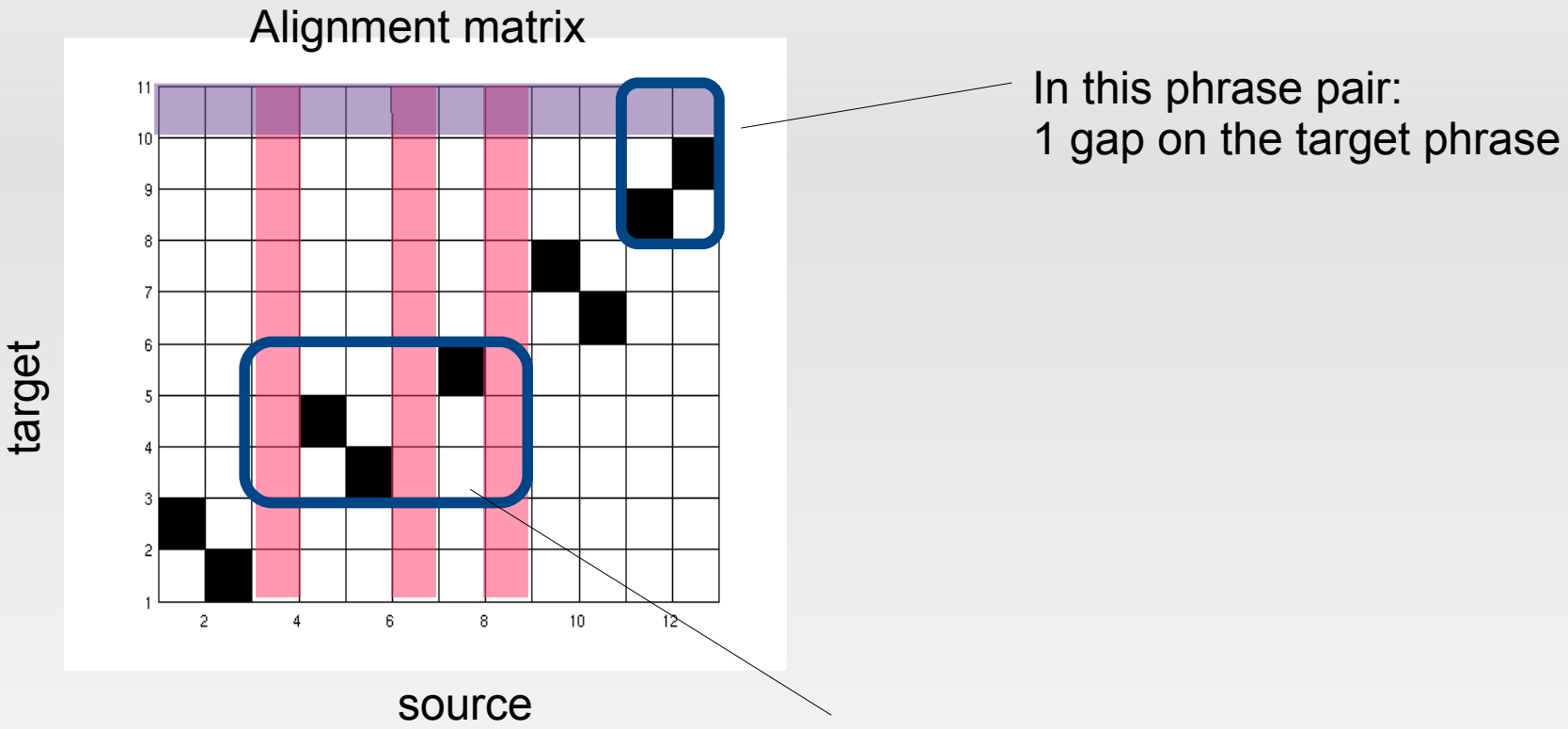
- Diversity of balance precision/recall between alignments.
- Usually precision prevails over recall.
- Two ways of describing to an alignment: links and unaligned words.
- In next section, we'll observe the importance of such factors in the generation of phrase pairs.

Analysis II: Phrase Extraction

Metrics

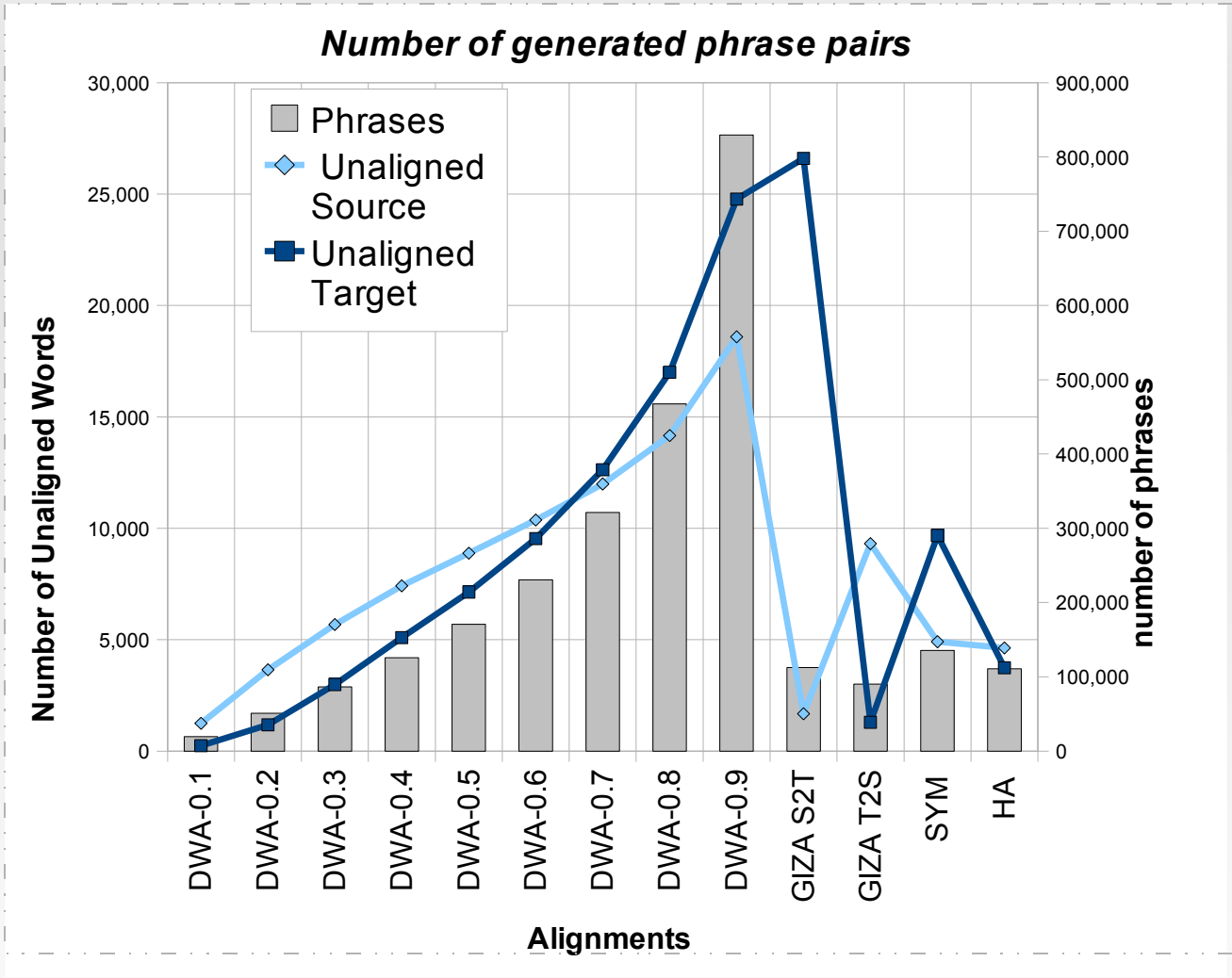
- Quantitative:
  - Number of phrases
  - Singletons (unique entries)
  - Phrase lengths
  - Gaps (unaligned words inside phrase pair)
- Qualitative:
  - Manual Evaluation

What do we mean by gaps?



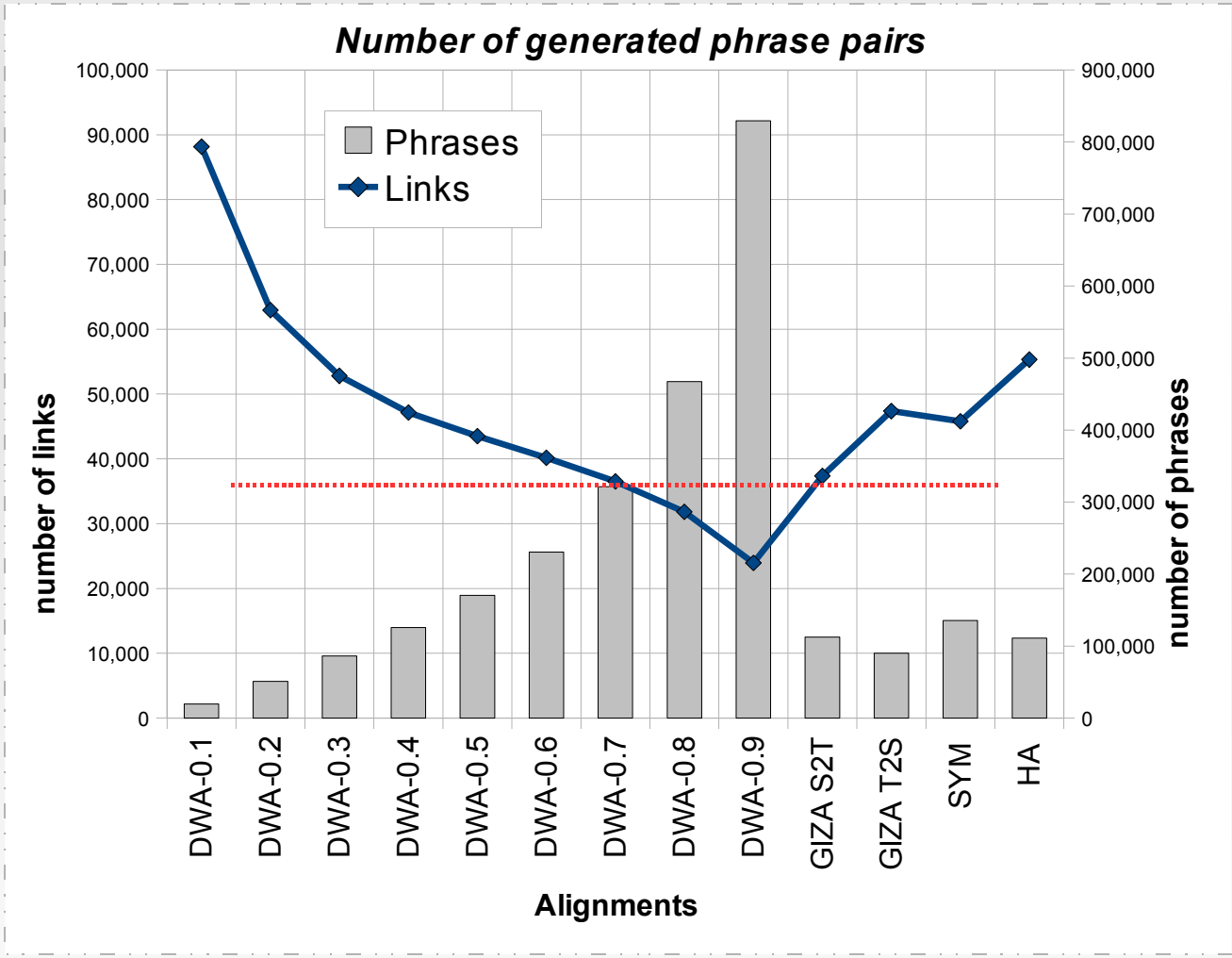
Number of Phrases

- PT **grows** as our alignment gets sparser
- Related to unaligned words rather than number of links



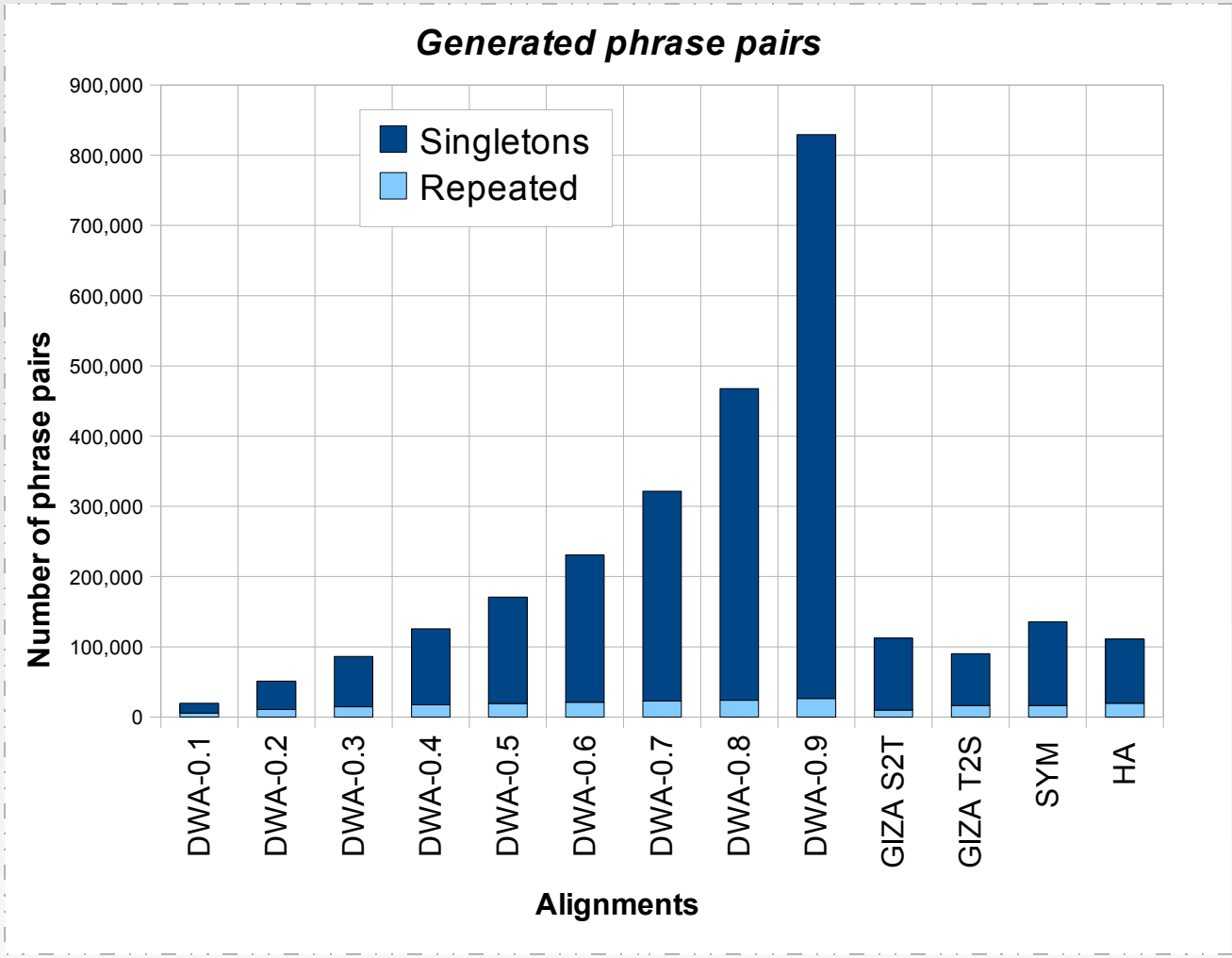
Number of Phrases

- PT **grows** as our alignment gets sparser
- Related to unaligned words rather than number of links



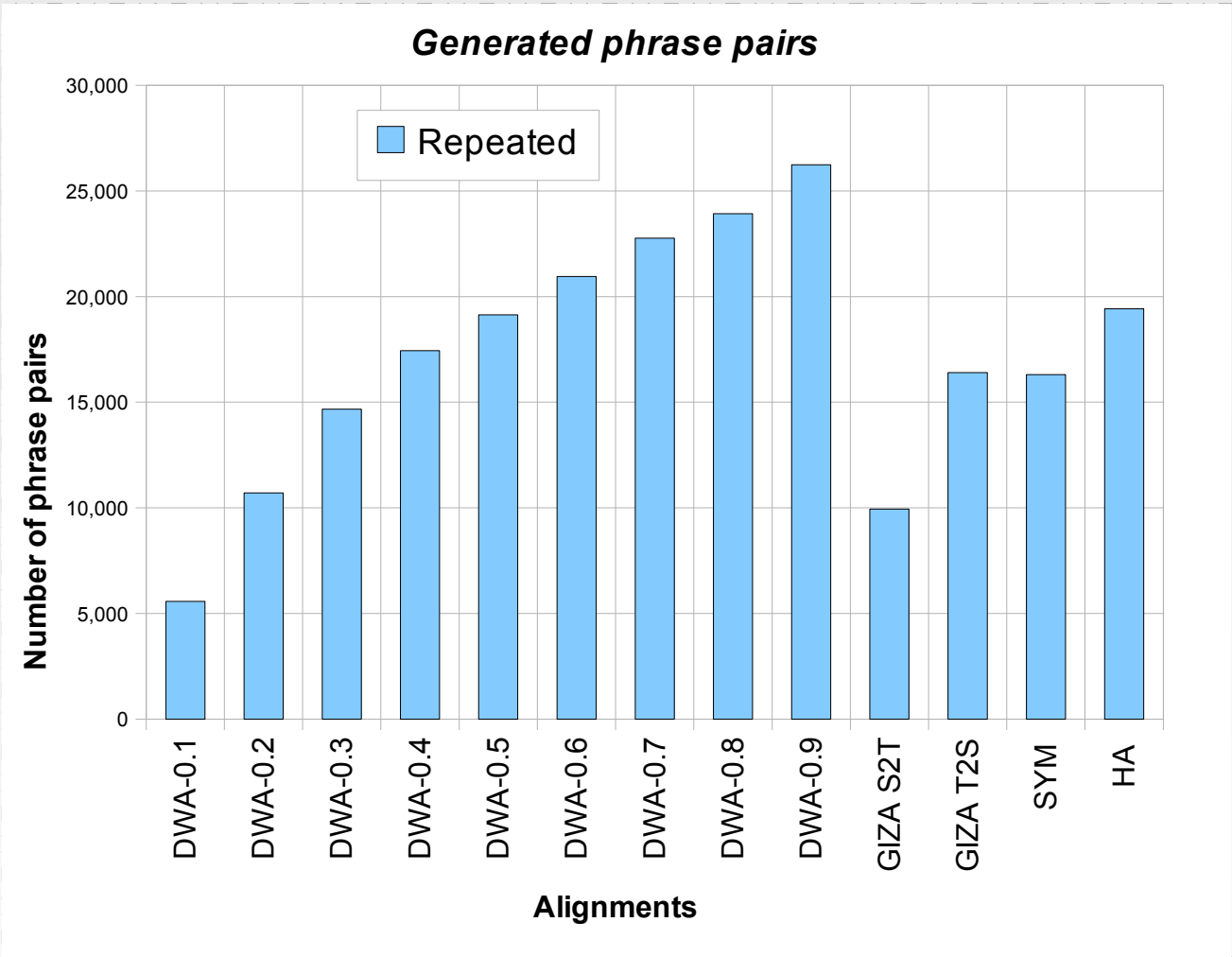
Singletons

- Most of the phrase-pairs are **singletons**.
- Repeated phrase-pairs grow at a slower rate.



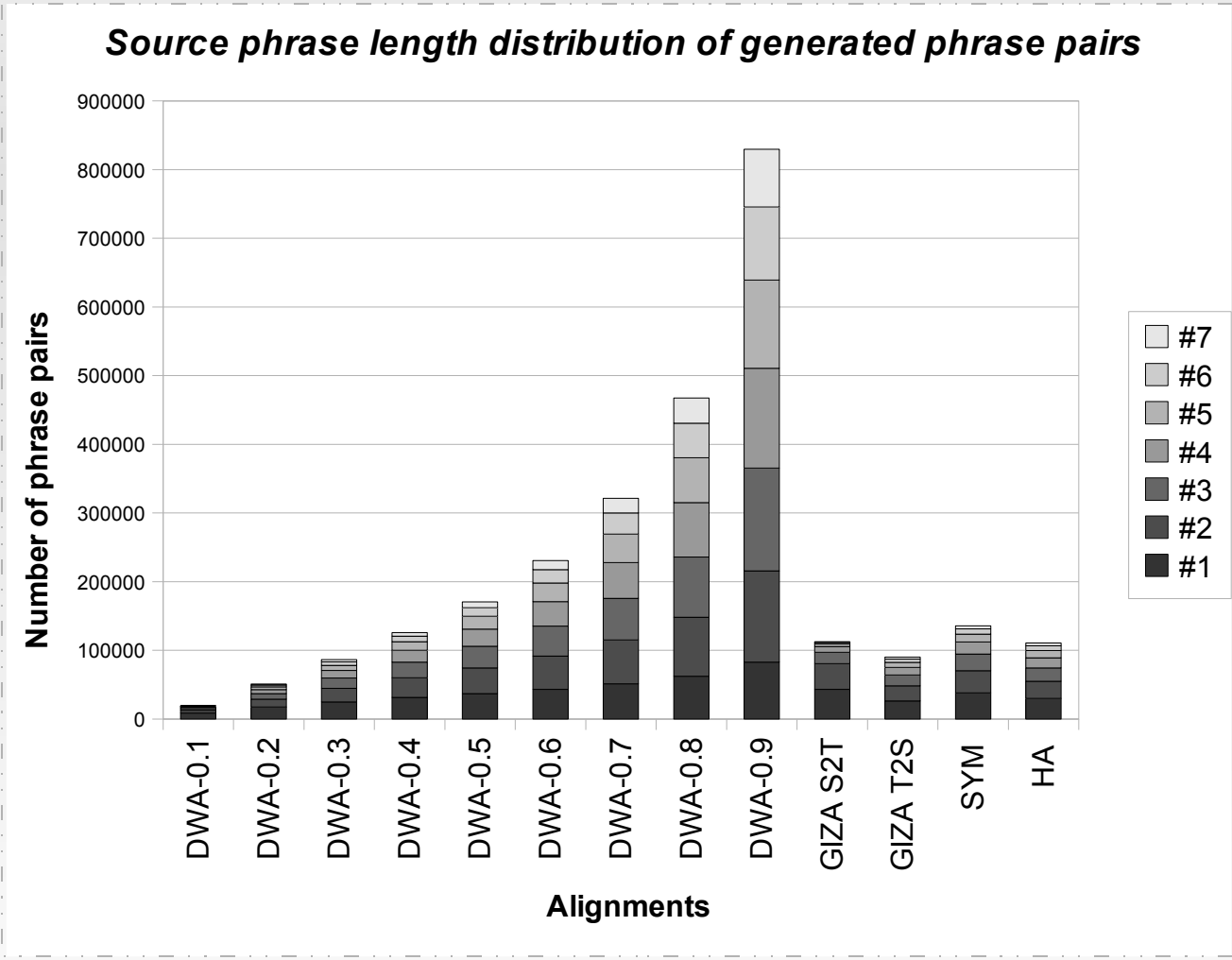
Singletons

- Most of the phrase-pairs are **singletons**.
- Repeated phrase-pairs grow at a slower rate.



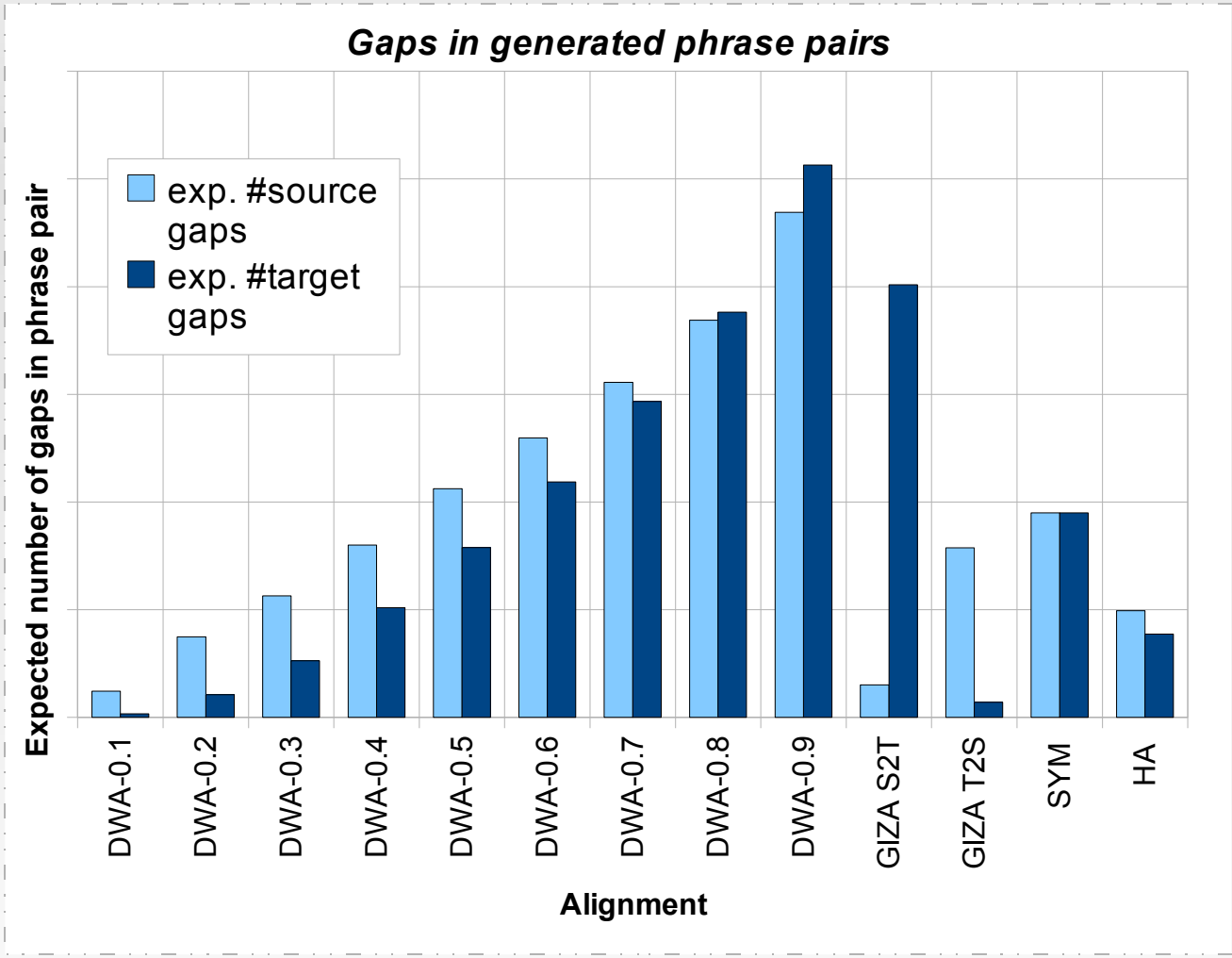
Phase Length

- As our PT grows, entries become longer and longer.



Gaps

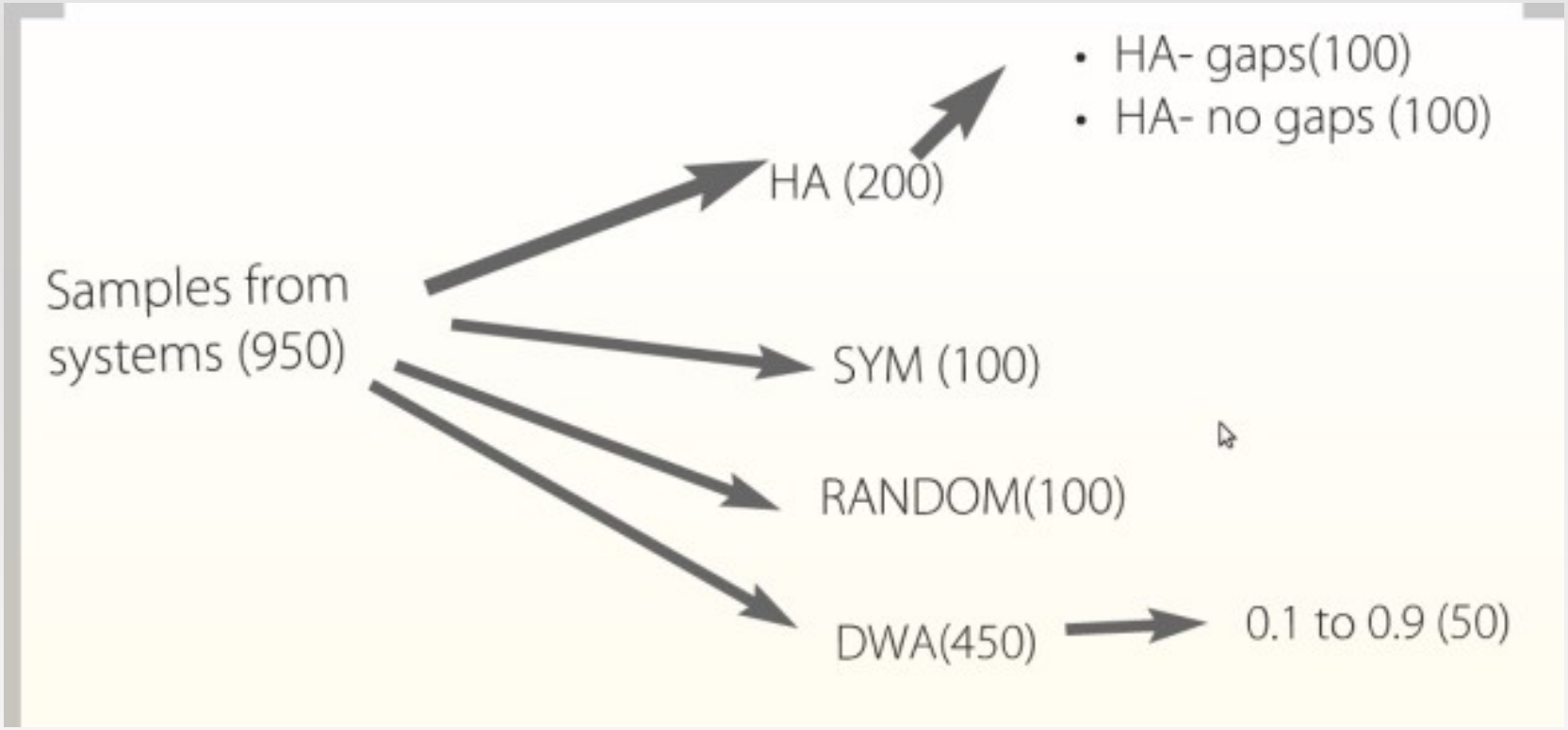
- The gaps inside a phrase pair increases too.
- The distribution of gaps in the generated phrases follows the distribution of unaligned words in the alignment



Human Evaluation of Phrase Pairs

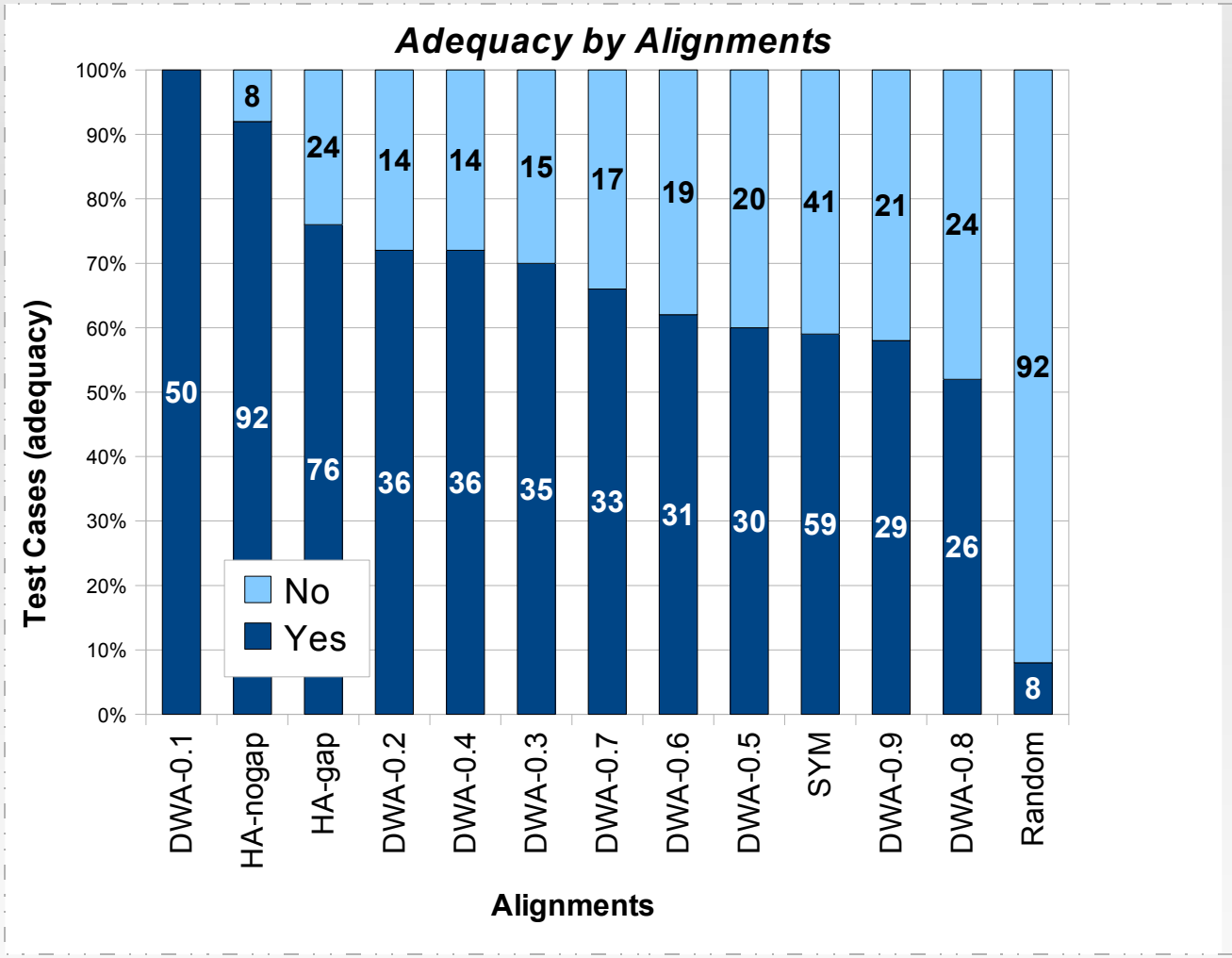
- Setup:
  - Native Chinese Speakers
  - Each subject was asked whether a phrase pair was adequate
  - No contextual information
  - Included a noisy input

Sampling

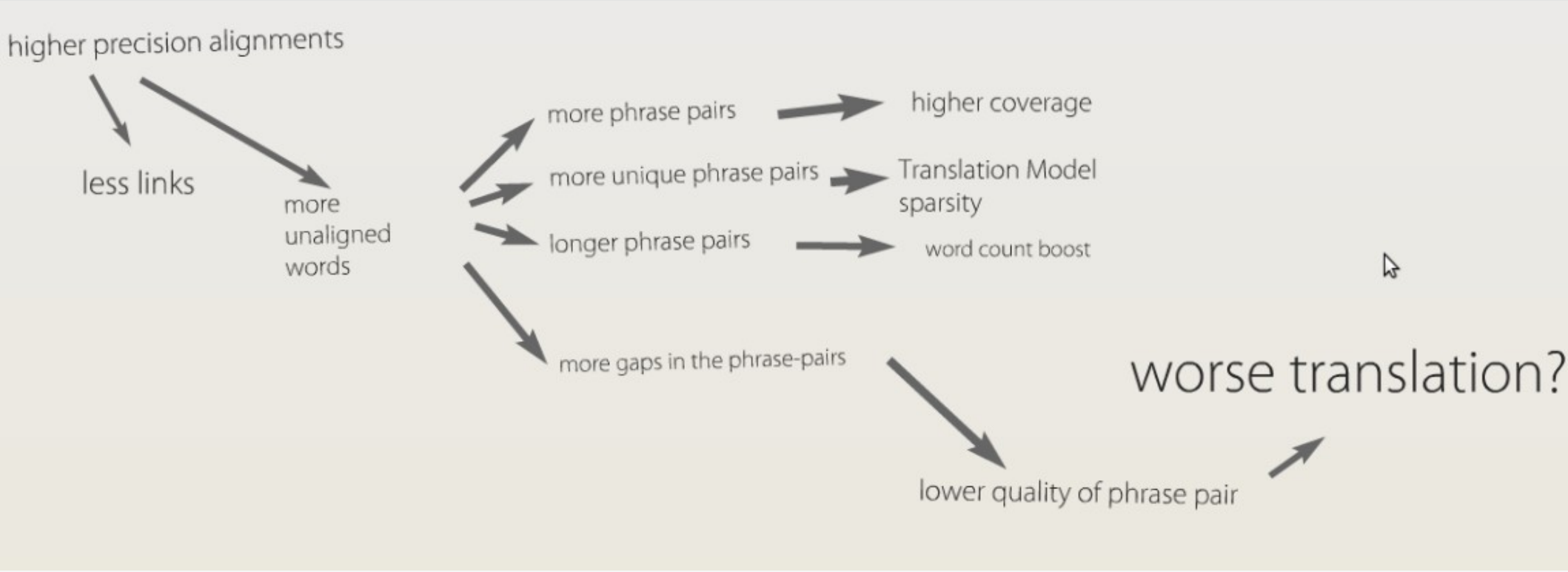

Samples from systems (950)
HA (200)
• HA- gaps(100)
• HA- no gaps (100)
SYM (100)
RANDOM(100)
DWA(450)
0.1 to 0.9 (50)

Results

- HA w/o gaps yields better results than HA w/ gaps.
- DWA-0.1 very good (short phrase pairs)
- DWA-0.5 not so great
- Random pairings are usually bad



Phrase Extraction: Summary



Lessons Learned: Mind your gaps

Taking into account GAPS

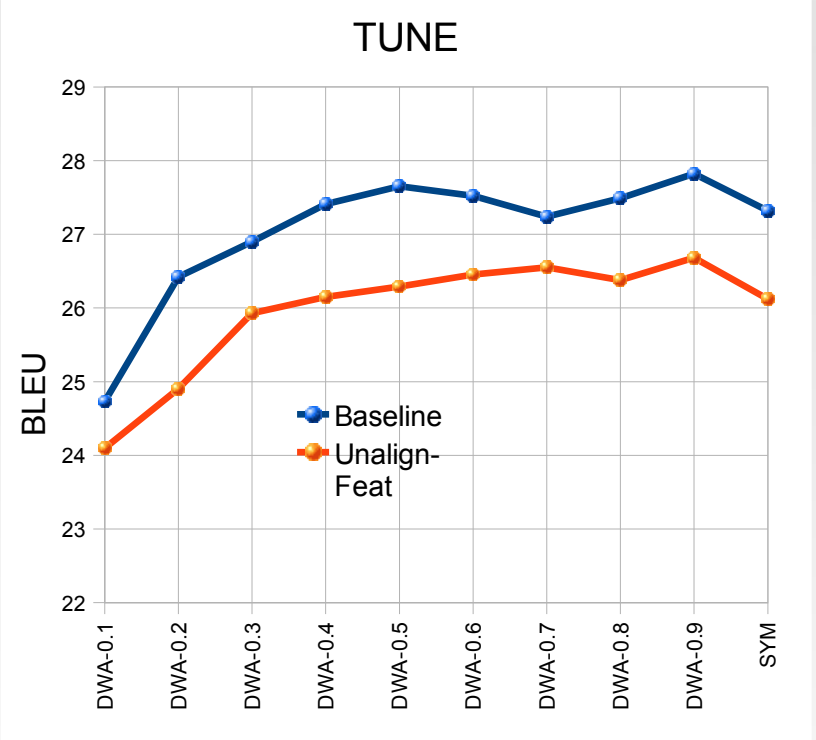
- Gaps inside phrase pairs have considerable impact on human perceived quality of phrase pair.
- Do they affect translation?
- Translation Experiment:
  - Include gap count as a feature (similar to WC)
  - Compare the performance of the different systems w/o the features

Setup

- Training
  - GALE P3 Data
  - Maximum sentence length 30
  - 1 Million sentences (random)
- Tuning
  - MT05
- Test
  - GALE DEV07-Blind

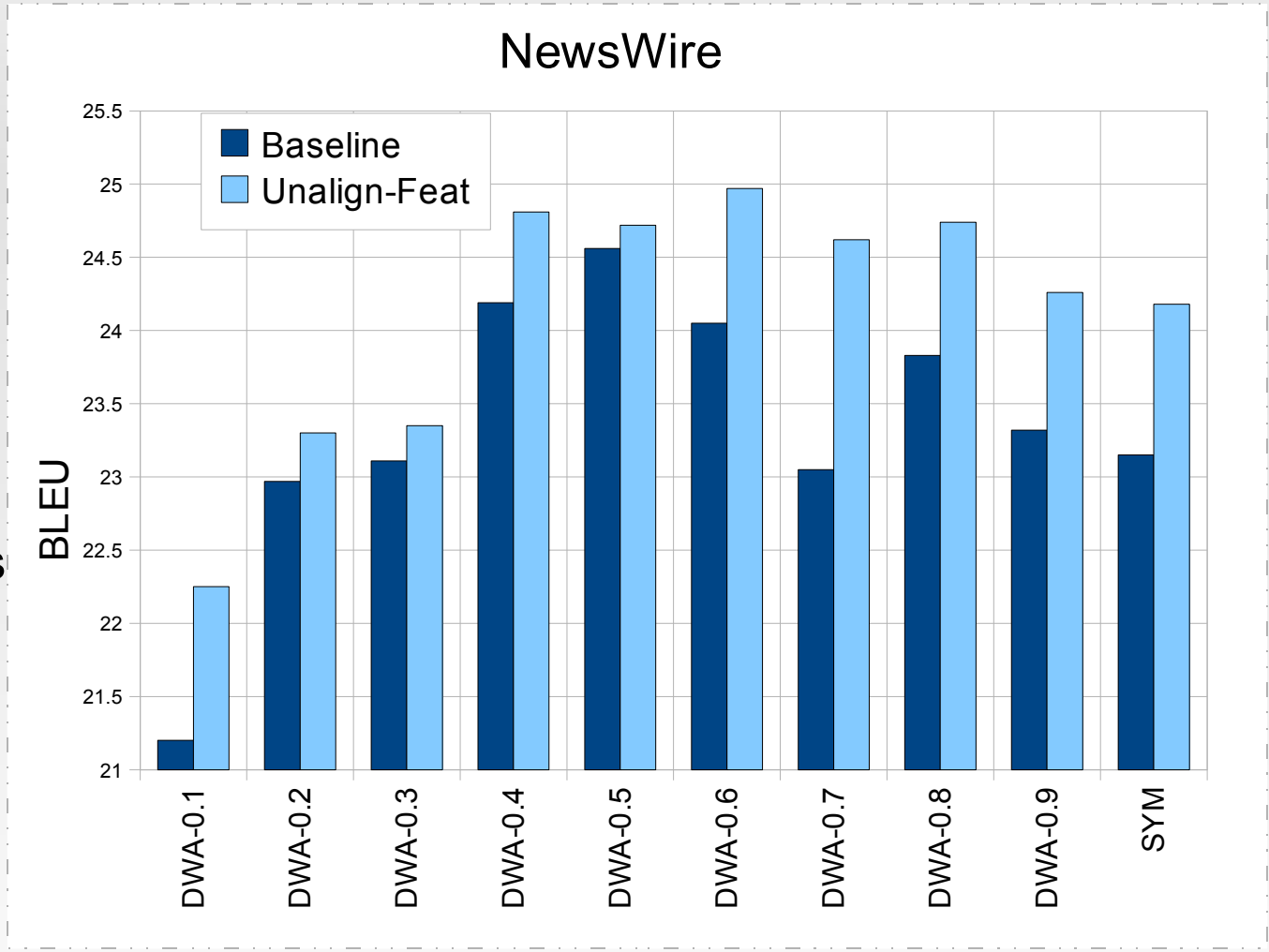
Tuning

- Baseline gets get better results
- Over-fitting?



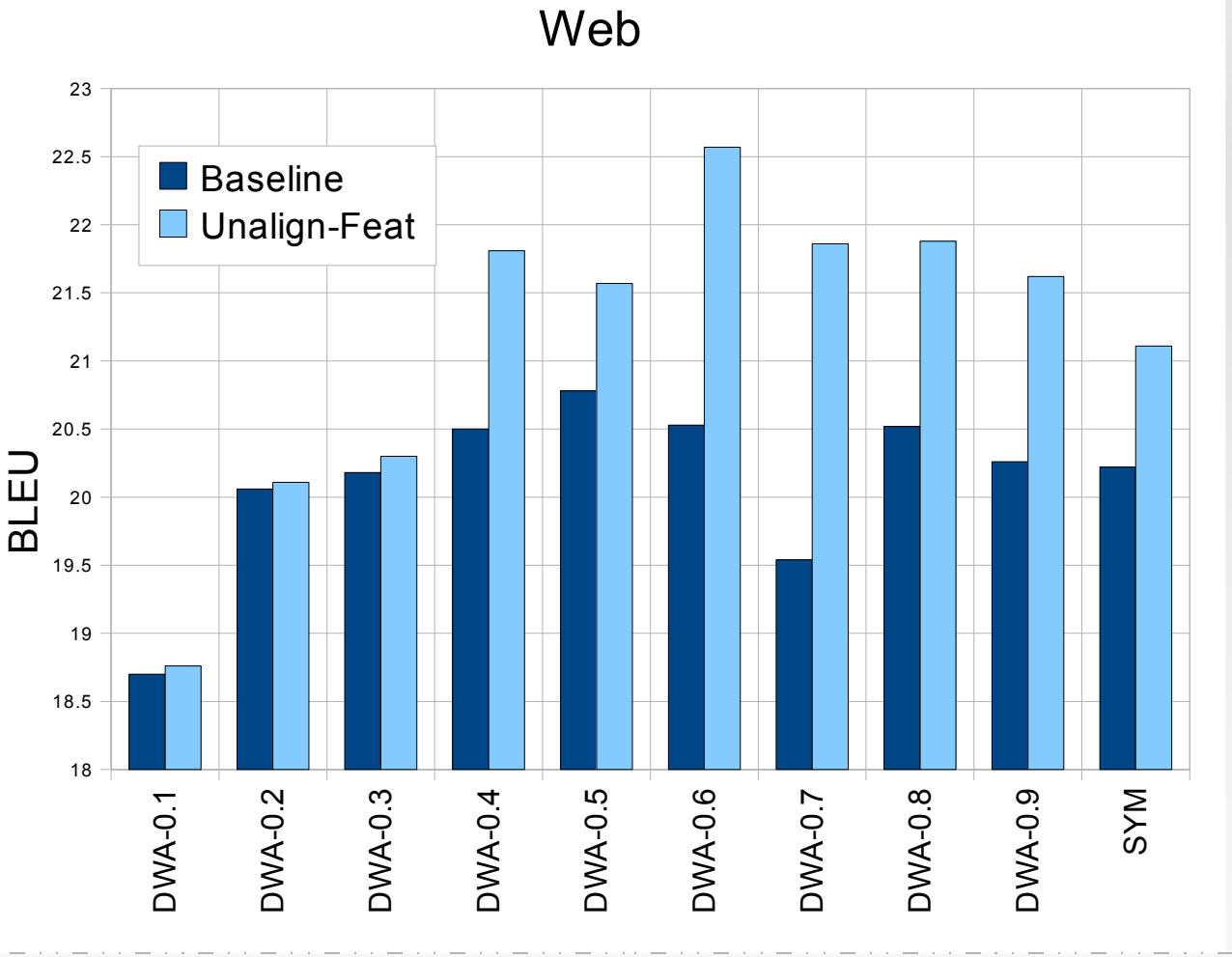
Experimental Results

- Overall gains
- Web performs better (~2 BP)
- Best system shifts to a higher precision alignment (DWA-0.5 => DWA-0.6)
- Higher recall alignments without much change



Experimental Results

- Overall gains
- Web performs better (~2 BP)
- Best system shifts to a higher precision alignment (DWA-0.5 => DWA-0.6)
- Higher recall alignments without much change



Conclusions

- We can describe an alignment by its quality and its structure (links, unaligned words).
- Unaligned words have an important role in phrase extraction (more than number links).
- The distribution of the gaps inside a phrase pair is related to the distribution of unaligned words in the alignment.
- Extracted phrase pairs with more gaps have lower human perceived quality.
- Taking into account the number of gaps in an extracted phrase pair as features achieved overall improvements.

Questions?

Adequacy of Phrase Pair

- by adequate we mean that a source phrase could be used as a translation of a target phrase in at least ONE situation, without loss of meaning

Experiment Variables

- Dependent:
  - Adequacy (yes/no)
- Independent:
  - System :
    - HA-Gaps
    - HA-No Gaps
    - DWA {0.1 ..0.9}
    - SYM
    - Random (noisy)
- Random
  - Number of Source, Target gaps
  - Evaluator

Results

- ANOVA
  - **System** is significative
  - Interaction **SourceGaps*TargetGaps** is significative
  - **Evaluator*System** is not significative