

Using Percolated Dependencies for Phrase Extraction in SMT

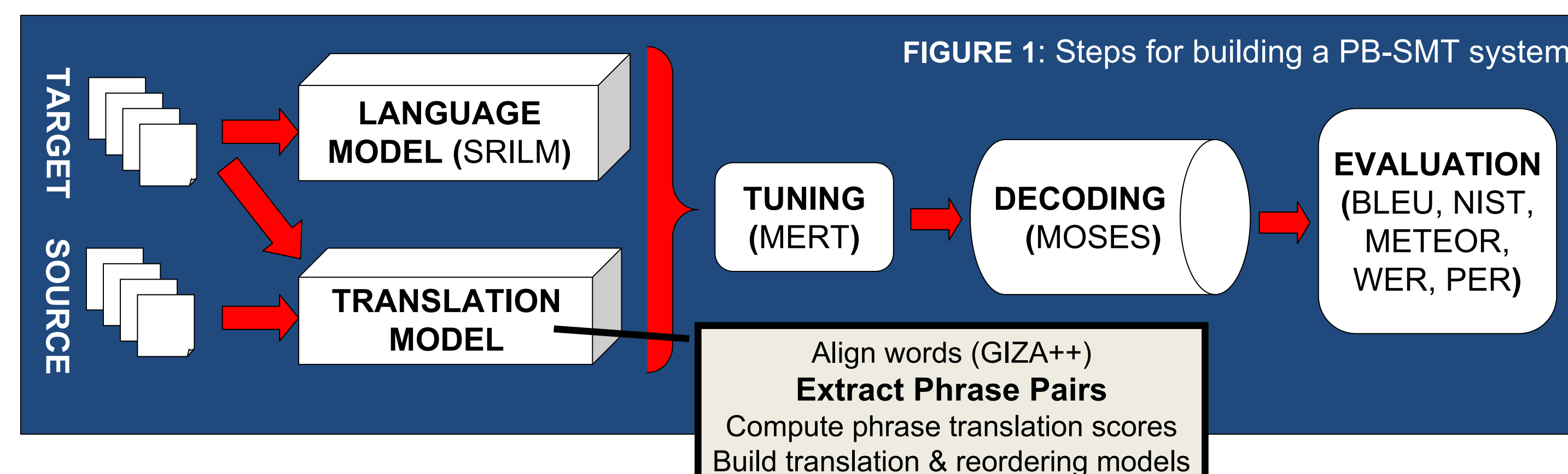
Ankit K. Srivastava

Andy Way

CNGL, School of Computing, Dublin City University, Ireland

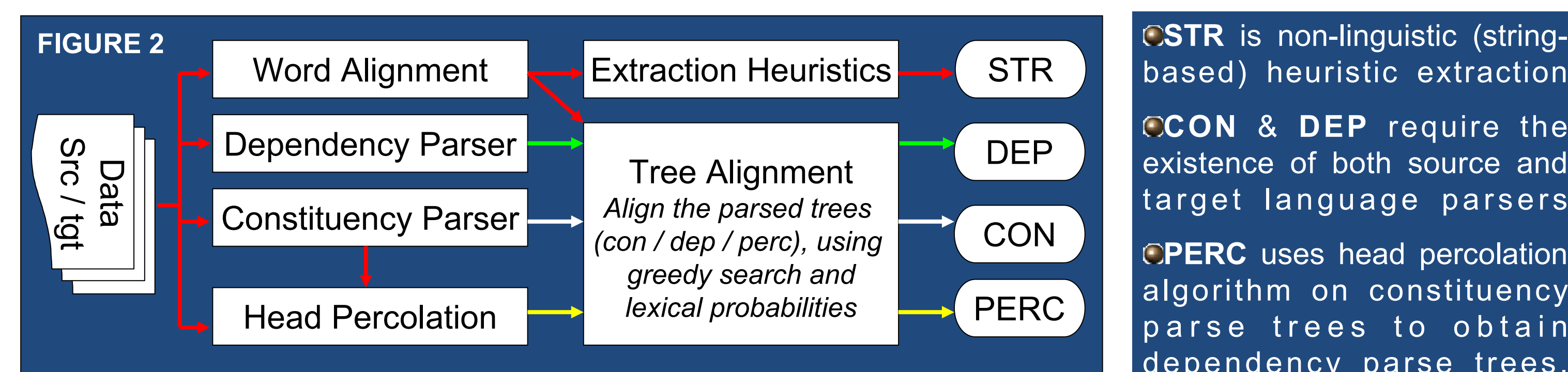
1. Introduction

There are numerous ways to extract phrase pairs in a Phrase-based Statistical Machine Translation (PB-SMT) system. Work to date has demonstrated the usefulness of phrase pairs induced from parallel treebanks (both **constituency** and **dependency**) in addition to the standard **n-gram** heuristics. We introduce phrase pairs induced from **percolated** dependencies (dependency structures obtained without using a dependency parser) as a unique knowledge source in the PB-SMT framework. We compare these four phrase extractions and perform experiments on 15 possible MT systems. Each system follows the same PB-SMT design (Figure 1) and only differs in the phrase extraction step of training translation models.



2. Four Phrase Extractions

Figure 2 demonstrates the modules used in our phrase extraction. Thus four types of phrase pairs are obtained: n-gram heuristically learned chunks (**STR**), constituency-parsed chunks (**CON**), dependency-parsed chunks (**DEP**), and percolated dependency-parsed chunks (**PERC**).



A head percolation table consists of hand-coded rules identifying the head-child of each node. It is used to select the head node (word) in each constituent structure which is then percolated like features to its parent projection. The algorithm outputs unlabelled dependency relations.

3. Experimental Analysis

- Experiments are carried out on both JOC (7,723 sentences) and a subset of the Europarl (100,000 sentences) English-French corpora
- 15 systems possible from the four phrase tables (S/C/D/P). Two or more tables are combined by merging and re-estimating probabilities. Results in Table 1
- Figure 3 shows how many phrase pairs are common between any two phrase tables (Europarl). The four blocks towards the right give the total number in each
- Figure 4 shows break-down of number of phrases the decoder used in decoding 2,000 Europarl sentences. (Moses run in 'trace' mode: S+C+D+P system)

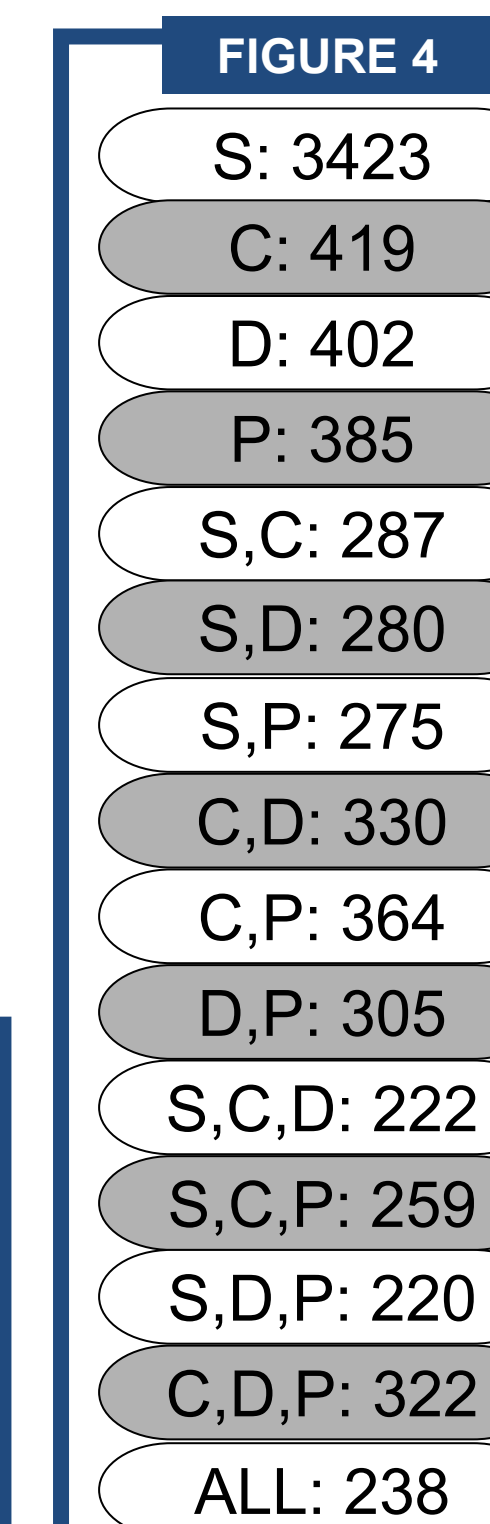
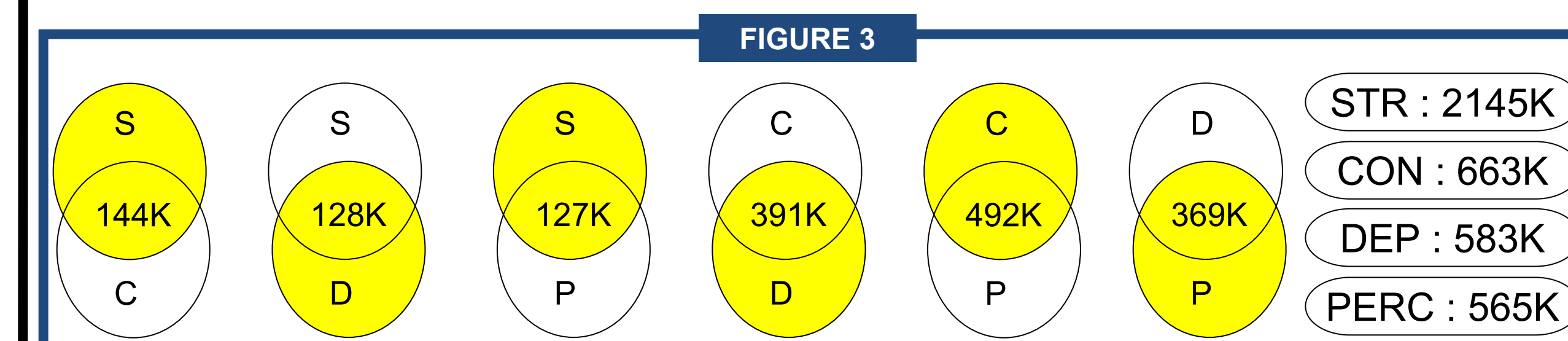


TABLE 1: Summary of results on Europarl Fr-En test data

SYSTEM	BLEU	NIST	METEOR	WER	PER
STR (S)	28.50	7.00	57.83	57.43	44.11
CON (C)	25.64	6.55	55.26	60.77	46.82
DEP (D)	25.24	6.59	54.65	60.73	46.51
PERC (P)	25.87	6.59	55.63	60.76	46.48
S + C	29.50	7.10	58.55	56.62	43.40
S + D	29.30	7.08	58.43	56.84	43.62
S + P	29.45	7.10	58.54	56.73	43.43
C + D	26.32	6.69	55.56	59.97	45.90
C + P	26.37	6.62	56.05	60.41	46.40
D + P	26.57	6.74	55.83	59.53	45.62
S + C + D	29.29	7.09	58.48	56.70	43.41
S + C + P	29.49	7.10	58.50	56.59	43.45
S + D + P	29.39	7.09	58.49	56.80	43.65
C + D + P	26.90	6.75	56.14	59.38	45.53
S+C+D+P	29.40	7.09	58.49	56.67	43.49

4. Conclusion

- Supplementing SMT (non-linguistic) phrases with syntax-aware phrases significantly improves translation accuracy
- PERC can successfully substitute DEP in the absence of one or more dependency parsers
- Each of the four phrase tables (S/C/D/P) has a considerable amount of uniqueness
- Adding PERC chunks to any system shows a general trend toward improving translation
- S+C' system performs best. However there is evidence that PERC chunks are still useful
- Future work includes discovering individual contribution of each phrase type and exploring better methods for combining phrase tables