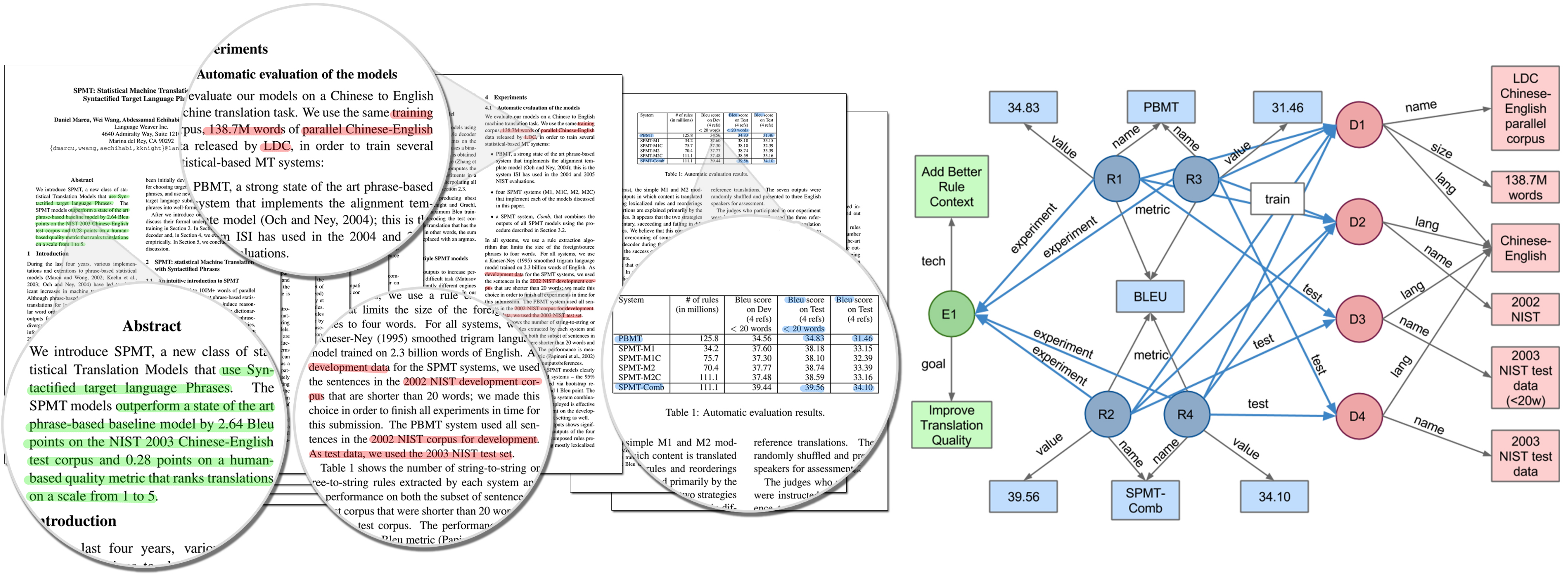
Extracting Structured Scholarly Information from the Machine Translation Literature

Eunsol Choi[†], Matic Horvat[‡], Jonathan May[*], Kevin Knight[*], Daniel Marcu[*]

[†]University of Washington, [‡]University of Cambridge, [*]University of Southern California Information Sciences Institute



System	# of rules (in millions)	Bleu score on Dev (4 refs) < 20 words	Bleu score on Test (4 refs) < 20 words	Bleu score on Test (4 refs)
PBMT	125.8	34.56	34.83	31.46
SPMT-M1	34.2	37.60	38.18	33.15
SPMT-M1C	75.7	37.30	38.10	32.39
SPMT-M2	70.4	37.77	38.74	33.39
SPMT-M2C	111.1	37.48	38.59	33.16
SPMT-Comb	111.1	39.44	39.56	34.10

Table 1: Automatic evaluation results.

Figure 1: An illustration of the extraction task. Starting with a scientific paper (left), information is extracted to construct a structured representation of the experimental data (right).

Motivation

Current technologies for searching scientific litearture do not support answering many queries that could significantly improve the day-to-day activities of a researcher. For instance, a Machine Translation (MT) researcher would like to answer questions such as:

- Which are the best published results reported on the NIST-09 Chinese dataset?
- What are the most important methods for speeding up phrase-based decoding?
- Are there papers showing that a neural translation model is better than a non-neural?

Current methods cannot yet infer the main elements of experiments reported in papers; there is no consenus on what the elements and the relations between them should be.

Structured Representation

We designed an MT specific representation of experimental information (shown in Figure 1, right), consisting of three interconnected intermediated representations. These representations in turn consist of **atoms**.

Datasets are corpora used to either to train or evaluate the systems. They consist of:

- name, e.g. `LDC Chinese-English parallel corpus`
- language, e.g. `Chinese-English`
- size, e.g. `138.7M words`

Results are experimental results presented in the paper, consisting of:

- numerical value, e.g. `34.83`
- metric, e.g. `BLEU`
- system name, e.g. `SPMT-Comb`

Experiment refers to the goal of the experiment and the method used to achieve it:

- goal, e.g. `Improve Translation Quality`
- technology, e.g. `Add Better Rule Context`

Data and Annotation

Annotation of 67 MT papers with structured representation was conducted by filling out a survey for each paper. Inter-annotator agreement on a subsample of 6 papers is shown in Figure 4.

To aid extraction, we processed each paper's PDF document using text and table extraction software to produce a structured text representation, split into sections and subsections with parsed tables accompanied by captions.

The dataset of 67 papers and annotations is available at:
https://github.com/isi-nlp/mt_lit_lrec16

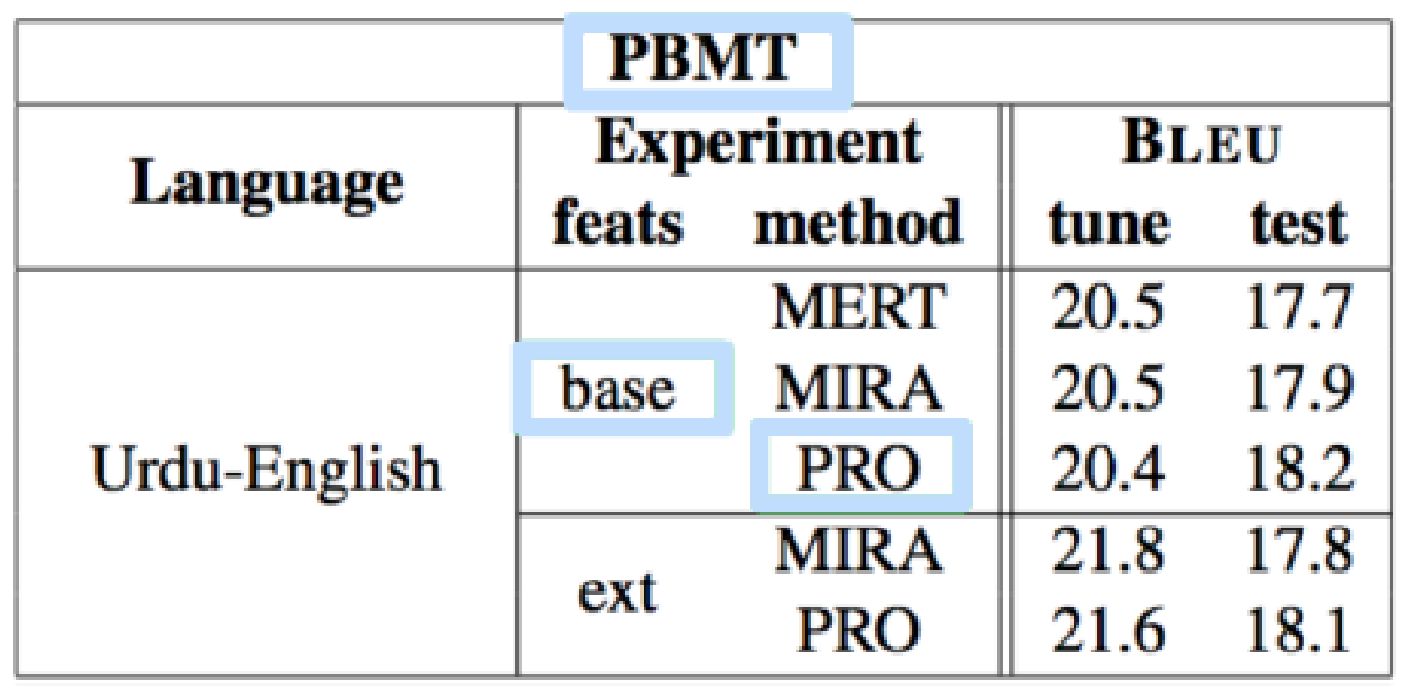
This work was partially sponsored by DARPA Big Mechanism (W911NF-14-1-0364).

Baseline System

We build a pipelined pattern-based system to serve as a baseline for the task. The system extracts individual atoms from a paper (Figure 1, left) and selects and links them into a structured representation (Figure 1, right).

Atom Detection

The aim of atom detection is to detect as many atoms as possible. They commonly use substring search, overlapping word search, and regex pattern matching in text or tables.



Language	PBMT Experiment feats	PBMT Experiment method	Bleu tune	Bleu test
Urdu-English		MERT	20.5	17.7
	base	MIRA	20.5	17.9
		PRO	20.4	18.2
	ext	MIRA	21.8	17.8
		PRO	21.6	18.1

Figure 2: Example table in an annotated paper (system name is '`PBMT base PRO`').

Linking

The aim of linking is to connect atoms into a structured representation using positional information of atoms. First, atoms are linked into intermediate structures representing datasets, experiment types, and results. Finally, intermediate structures are linked to form the full structured representation.

Evaluation

Evaluation is done on the set of 62 documents. Atom detection was evaluated in terms of recall, while overall performance was evaluated in terms of Smatch F1 score.

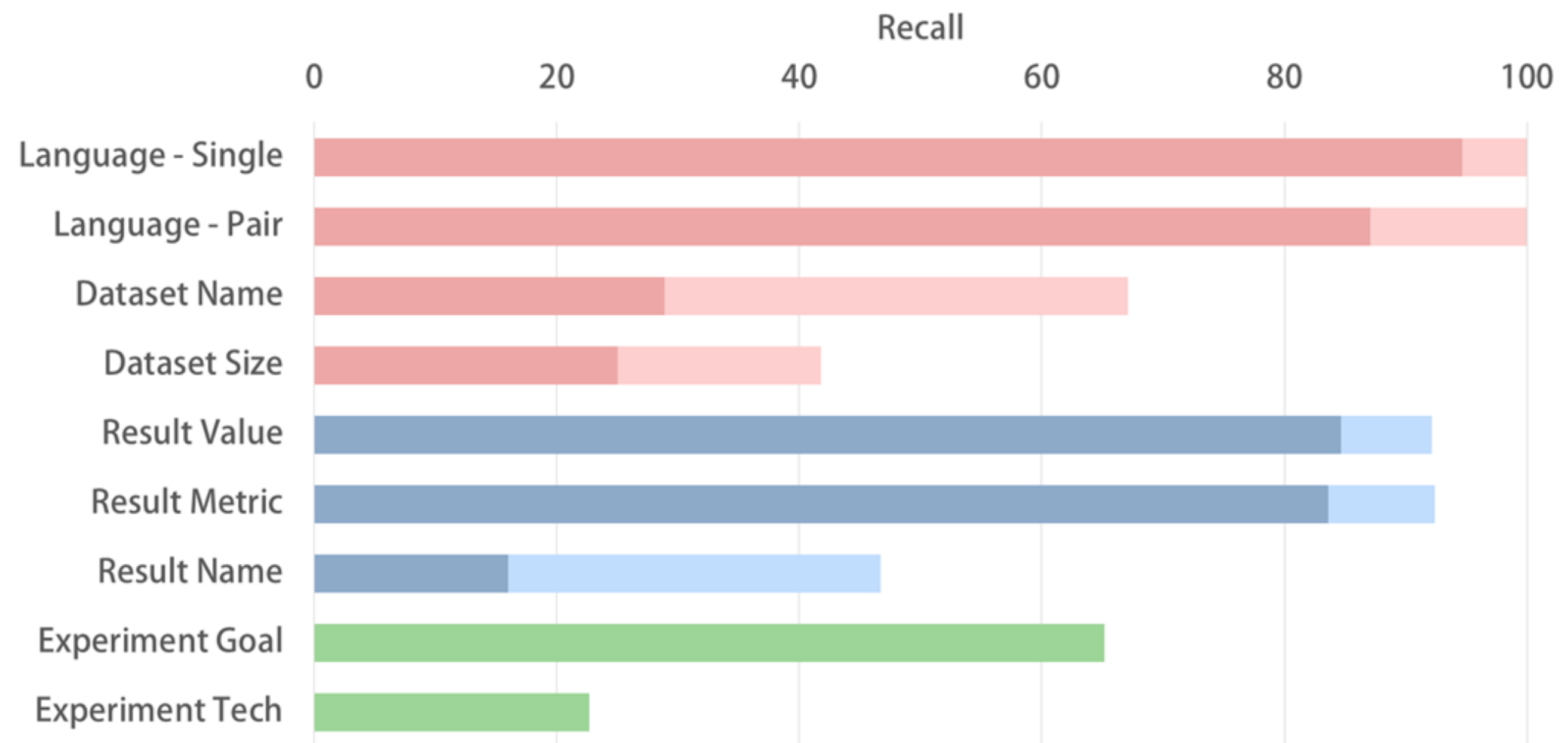


Figure 3: Detection performance in terms of Recall (light: reconstruction from surveys).

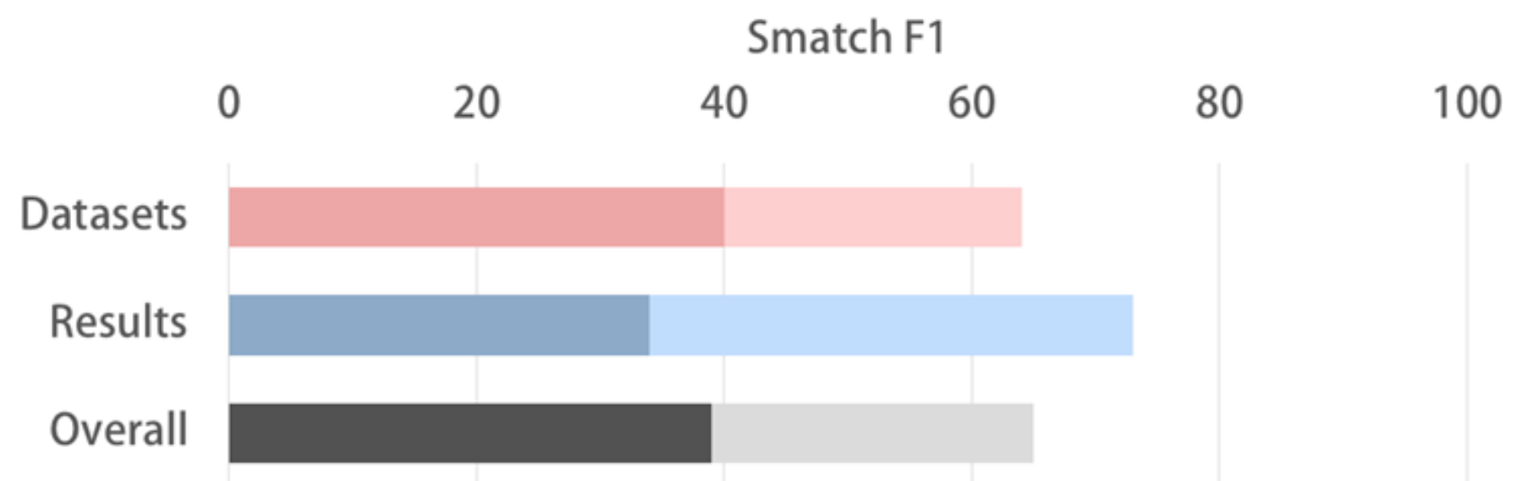


Figure 4: Linking performance in terms of Smatch F1 score (light: inter-annotator scores).