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, Information Sciences Institute[*]
eunsol@cs.washington.edu, matic.horvat@cl.cam.ac.uk, {may,knight,marcu}@isi.edu

Abstract

Understanding the experimental results of a scientific paper is crucial to understanding its contribution and to comparing it with related work. We introduce a structured, queryable representation for experimental results and a baseline system that automatically populates this representation. The representation can answer compositional questions such as: "Which are the best published results reported on the NIST 09 Chinese to English dataset?" and "What are the most important methods for speeding up phrase-based decoding?" Answering such questions usually involves lengthy literature surveys. Current machine reading for academic papers does not usually consider the actual experiments, but mostly focuses on understanding abstracts. We describe annotation work to create an initial ⟨scientific paper; experimental results representation⟩ corpus. The corpus is composed of 67 papers which were manually annotated with a structured representation of experimental results by domain experts. Additionally, we present a baseline algorithm that characterizes the difficulty of the inference task.

Keywords: Information Extraction, Scientific Literature, Structured Prediction

1. Introduction

Current technologies enable one to access large scientific literature repositories via a variety of means, which range from simple keyword searches for content and authors to sophisticated inferences that exploit citation links (Dunne et al., 2010; Schäfer et al., 2011), techniques that automatically identify sections and section labels (Teufel and Kan, 2011), and unsupervised methods to infer information structures (Kiela et al., 2015). Unfortunately, these access methods fall short of supporting many queries that could significantly improve the day-to-day activities of a researcher. Imagine, for example, a young researcher who wants to begin working on Machine Translation (MT) or a seasoned researcher who wants to keep track of recent developments in the field. Ideally, they would like to quickly get answers to questions like:

- Which are the best published results reported on the NIST-09 Chinese dataset?
- What are the papers that show on training sets larger than 100M words that morphology-inspired models lead to improvements in translation quality that are statistically significant?
- 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 model?

To our knowledge, answering such queries is beyond the state of the art. Current methods cannot yet infer the main elements of experiments reported in papers; as a matter of fact, no consensus exists on what these elements should be and what the relations between them are.

In this paper, we take a few steps towards addressing these shortcomings. By focusing on MT as our exemplary subfield of study, we propose a representation that explicitly models the hidden structure of typical experiments: data sources used for training and testing, evaluation metrics, languages, baseline algorithms, methods and algorithms that experiments are meant to highlight, etc. We also report annotation work aimed at creating a gold standard for this task, and we review a set of simple algorithms that we developed as a baseline to objectively characterize the difficulty of the task. By making our representations, data, and baseline results public, we hope to contribute to the more general effort of transitioning the Information Extraction field from identifying simple mentions and relations to identifying and reasoning with complex structures like events, scripts, and experiments.

[†] Work done during internship at ISI. Both authors contributed equally.

2. Structured Representation of MT Experiments and Task Definition

To capture meaningful elements of experiments in MT conference papers, we design a structural representation of experimental results. While this can be used as a reference to understand experiments in other fields, we intentionally designed it to answer meaningful queries about MT papers. Our overall task is to convert a paper (Figure 1, top) into a connected graph (bottom) of experimental results. Figure 1 shows an example of a paper "SPMT: Statistical Machine Translation with Syntactified Target Language Phrases" (Marcu et al., 2006). The structured representation is composed of DATASETS, EXPERIMENT TYPE, and RESULTS.

Datasets are corpora used to either to train or evaluate the systems. We decompose datasets into name, size, and language. The example uses four datasets, including `LDC Chinese-English parallel corpus` and `2002 NIST`. Only the first dataset has a stated size, `138.7M words`, while all of them use the `Chinese-English` language pair.

Experiment type refers to the goal of the experiment and the method used to achieve it. We define 9 goals and 27 methods.

Results are experimental results presented in the paper, consisting of numerical value, metric, and the name of the system that achieved the result. In Figure 1, we retrieved four values (`34.83`, `31.46`, `39.56`, `34.10`) with the `BLEU` metric.

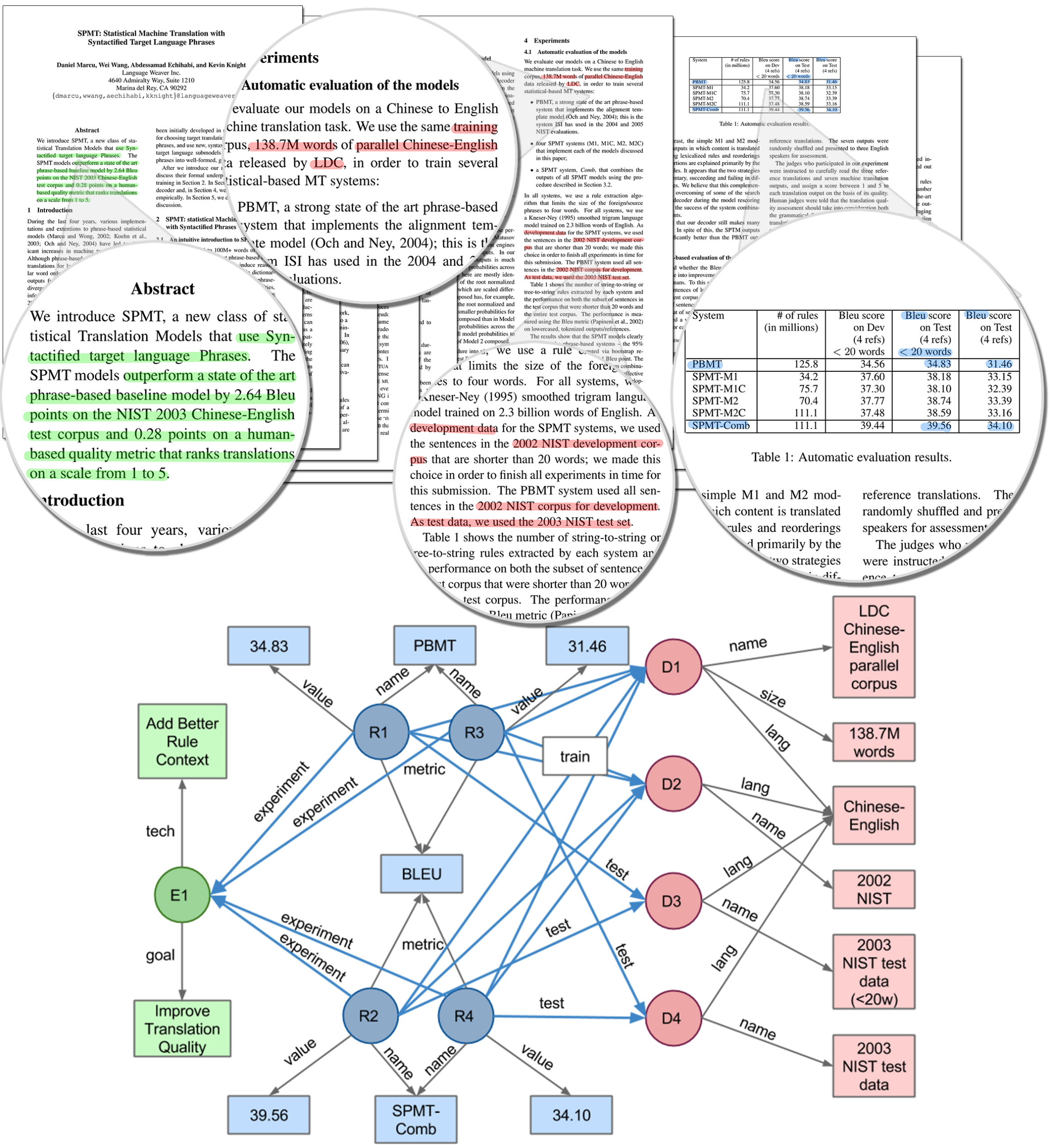

Figure 1: An illustration of the extraction task. Starting with a scientific paper (top), information is extracted to construct a structured representation of the experimental data (bottom).

Two result values are associated with the `PBMT` system and two with the `SPMT-Comb` system. The RESULT values are connected to DATASETS via the `test` or `train` relation, and to an EXPERIMENT TYPE.

We refer to the individual pieces of information that comprise the structured representation as *atoms*. In the example, there are 15 atoms, including `Chinese-English`, `BLEU`, `31.46` and `2002 NIST`.

3. Data and Annotation

In order to construct the dataset, we started by selecting papers related to Machine Translation from the ACL Anthology corpus (Radev et al., 2009) using keyword search and targeting of MT related workshops. For a random sample of 67 papers, we asked the annotators to provide a structured representation of experimental results as defined in the previous section. In order to aid structured information extraction, we automatically produced a structured text representation of each paper. The representation consists of plain text split into sections and subsections, as well as parsed tables. The annotated dataset is released alongside the paper[1] to promote future research. In total, 1063 atoms were annotated. Additional dataset statistics are presented in Table 1. In the remaining part of this section we describe the dataset construction and annotation in more detail.

3.1. PDF to Structured Text Conversion

The starting point is a set of papers in PDF format. We convert papers in PDF format into structured text using the

[1]The dataset is available from https://github.com/isi-nlp/mt_lit_lrec16.git

		Avg. Count	St. Dev.
Text	Sections	7.55	1.38
	Tables	3.76	1.78
	Sentences	272.55	89.65
Experiment Structure	Total Atoms	23.47	9.85
	Train/Tune Dataset	4.33	4.14
	Test Dataset	2.01	1.52
	Result	6.84	5.01
	Experiment Type	1.13	0.34

Table 1: Structured text and survey response mean and standard deviation.

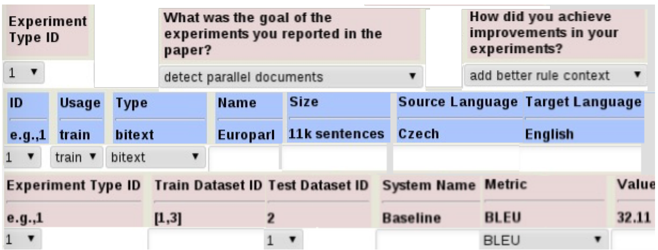

Figure 2: Survey form to collect annotations.

commercial TET system.[2] As tables are often used to report experimental results, we pay special attention to their extraction. We extract tabular information using TableSeer[3] (Liu et al., 2007). We use ParsCit[4] (Councill et al., 2008) to derive the hierarchical structure of sections and subsections. We produce the final representation of papers with a system that combines the inputs of all three components. The process produces structured text, split into sections and subsections with parsed tables accompanied by captions, but does not include figures.

3.2. Structured Representation Annotation

Annotators are presented with the papers in PDF format. Ideally, annotators would highlight relevant information in the text and link it to the structured representation. However, such linking is very time consuming. As an alternative, we design a survey annotation tool, shown in Figure 2. From survey responses, we create the structured representation deterministically. We gather the annotations by sending the survey to the selected papers' authors and by annotating them ourselves. From survey responses, we create the structured representation deterministically.

Six papers were annotated by two annotators. Inter-annotator agreement on these papers is shown in Table 3. Annotators disagree frequently on techniques, as a single paper can use multiple techniques. Lexical variability in naming a dataset or a system also causes disagreement. Instructed to choose the top and baseline performance for each important evaluation, people at times chose different experiments.

[2]PDFlib TET 4.4 Text Extraction Toolkit
[3]http://sourceforge.net/projects/tableseer
[4]http://aye.comp.nus.edu.sg/parsCit

4. Baseline System Approach

We present a pipelined pattern-based system that extracts individual atoms from a plain text logical representation of a machine translation paper and selects and links them into a structured representation.

4.1. Atom Detection

In atom detection, the system generates lists of candidates for each atom type. The aim of atom detection is to detect as many atoms as possible to enable subsequent steps in the system to select among multiple candidates. Detection consists of finding substrings, overlapping words, and matching regex patterns in text or tables.

The **language detector** matches a pre-defined list of languages against the text. The list includes two and three-character language abbreviations.

The **dataset size detector** is based on regex pattern matching expressions such as '8M sentence pairs.' These patterns either include a unit (as above) or are unitless, e.g., '8M.'

The **dataset name detector** matches a curated list of known MT datasets to text. Various ways to express datasets are encoded by regex patterns.

The **system name detector** finds candidates in result tables, excluding numerals and specific keywords.

The **result value detector** captures numeric cells in result tables, such as 24.3, 12%.

The **result metric detector** is based on a list of common metrics used in MT such as BLEU.

The **goal detector** and **technology detector** match pre-constructed lists of phrases.

4.2. Linking

Linking consists of two stages: (1) linking of atoms into intermediate structures and (2) linking of intermediate structures into the final composite structure. In the first stage, individual atoms are selected and linked together to first form a structure representing either a DATASET, an EXPERIMENT TYPE, or a RESULT. At this stage, many atoms are available to link and a selection process is carried out on atoms to create candidate intermediate structures. In the second stage, a selection process is subsequently carried out on intermediate structures to create the final composite structure representing a single paper.

Dataset We select language pairs based on frequency, and we find the closest dataset name and size atom. We choose edge labels by searching for keywords such as 'train' and 'test' in proximity.

Result We construct RESULT structures from tables, using the column, row, and caption of tables. We link system name atoms and result metric atoms found either in the first row or the first column.

Composite DATASETS are linked to RESULTS based on proximity measures and cues from text, for example mentions of languages or dataset names in captions or adjacent table cells. We limit each result to a single test DATASET, but allow multiple training DATASETS.

Atom Type		P	R	F1	R*
Language	Single	21.3	**94.7**	34.8	100.0
	Pair	76.3	**87.1**	81.3	100.0
Dataset	Name	20.1	**28.9**	23.7	67.1
	Size	24.3	**25.0**	24.6	41.8
Result	Value	7.1	**84.7**	13.0	92.2
	Metric	35.5	**83.6**	49.8	92.4
	Name	7.1	**16.0**	9.9	46.7
Experiment	Goal	72.6	**65.2**	68.7	-
	Tech	24.2	**22.7**	23.4	-

Table 2: Performance of atom detector in terms of Precision, Recall, and F1 score, and reconstruction from survey response (R*).

PBMT				
Language	**Experiment**		**BLEU**	
	feats	**method**	**tune**	**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 3: System name annotated as '`PBMT base PRO`'.

		P	R	F1
Baseline	Atom	0.35	0.18	0.22
	S-Dataset	0.51	0.40	0.40
	S-Result	0.54	0.31	0.34
	S-Total	0.58	0.34	0.39
Inter Annotator	Atom	0.44 (Jaccard Score)		
	S-Dataset	0.66	0.66	0.64
	S-Result	0.77	0.77	0.73
	S-Total	0.68	0.68	0.65

Table 3: Linking performance evaluation results in terms of Precision, Recall, and F1 Smatch scores. In the case of atoms, precision, recall, and F1 of the selected atoms after linking is shown.

5. Evaluation

Data From the collected data, five papers are used for development, and 62 are used for evaluation.

Evaluation Metrics We evaluate system performance of atom detection with precision and recall. We approach the evaluation of the linked structured representation by transforming it into a directed acyclic graph and computing the Smatch score (Cai and Knight, 2013), previously used to evaluate the similarity between Abstract Meaning Representation (AMR) structures.

5.1. Atom Detection Evaluation

Table 2 shows the performance of atom detection. As annotators do not tell us where the information is located, we match the annotated atoms to every substring in the structured text and present annotation recall from text as R* in Table 2. This presents a soft ceiling for our baseline approach. Finding annotated dataset name, size, and system name atoms was challenging due to abbreviations, PDF-to-text conversion errors, lexical diversity and name expansion, as well as extracted values consisting of scattered strings, as shown in Figure 3. These problems persist in atom detection.

Successes The language and language pair detectors achieve high recall. The result value and metric detectors also achieve high recall on par with the R*. Dataset name and size detectors achieve around half of the gold recall. Even when they do not extract the entire name correctly, they often capture a substring. Dataset names and sizes can be expressed in a variety of ways—correct system output can differ from the annotation.

Unresolved challenges The language pair detector struggles with phrases such as 'translating English to Japanese and Turkish,' detecting only 'English-Japanese.' This name expansion happens in dataset names as well. Correctly interpreting information in tables, where information is presented in a structured manner analogous to Figure 3 is also interesting problem to be resolved. For datasest, errors are primarily due to the variety of ways to express dataset names and sizes (for example, `MT03` – `MT08` refers to a set of 6 datasets). Additionally, new unnamed datasets are constantly being introduced into the literature. Detecting system names is even more challenging, as there is no naming convention. Detecting the goal and technology of a paper achieves mediocre recall. This challenging problem is a focus of research by itself (e.g. Gupta and Manning (2011)).

5.2. Linker Evaluation

The linking performance is presented in Table 3. As a reference point, we present scores computed between two gold annotations as 'Inter-Annotator.' Analysis shows that the system can detect the highest scoring BLEU score on the correct language pair, but at times fails to recover the names of datasets or systems. For instance, in the example in Figure 1, the system is able to detect all four result values and the name `PBMT`, and link correctly to the language pair and metric, but it could not retrieve the name `SPMT_Combo`. Furthermore, correctly linking a result to a set of training corpora is a challenge that can only be resolved by understanding long-distance dependencies in the document. While it is able to link subsets of datasets, the system often fails to recover the full expanded names of datasets. For the example in Figure 1, the system is able to retrieve `NIST` as both test and training data, but without specifying the year.

6. Related Work

Automatically processing scientific literature is receiving growing attention. Researchers focus on extracting information from abstracts, titles, and citations. There have been efforts to create extractive summaries (Abu-Jbara and Radev, 2011; Qazvinian et al., 2013) and flows of scientific ideas (Shahaf et al., 2012). Analysis of individual papers (Tsai et al., 2013; Gupta and Manning, 2011; Kiela et al., 2015) focuses mainly on abstracts.

J. Hutchins has manually compiled an electronic repository of machine translation literature.[5] He categorizes 11,500 papers by methodology, language pairs, systems, linguistic aspects, etc.

For the applications of the automatic analyses, the iOpener project (Dunne et al., 2012; Dunne et al., 2010) and Schafer et al. (2011) present bibliometric lexical link mining, summarization techniques, and visualization tools. These focus on metadata such as keyword, author, institution, conference name, and citations.

7. Conclusions and Future Work

Presenting experimental information from scientific papers in a structured representation that supports queries will help researchers in understanding scientific literature. To this end, we propose a new task of automatically extracting experimental information from scientific papers. We focus on the field of Machine Translation, for which we created a structured representation capturing the experimental information. We create a dataset of 67 MT papers with manually annotated experimental information in the structured representation. The dataset is available at https://github.com/eunsol/mt_lit_lrec16.git. Finally, we evaluate a simple baseline system, which demonstrates several challenges for automatic extraction of experimental information. These include finding and resolving structured information in tables, dealing with lexical variability and resolving long distance connections. Future work can explore injecting domain knowledge in the form of prior beliefs such as result ranges for metrics, as well as using manually compiled repositories for distant supervision.

Acknowledgements

We thank reviewers for helpful feedback. Our gratitude goes to our annotators, including authors of the machine translation papers who replied to our survey and members of the ISI natural language processing group. This work was partially sponsored by the DARPA Big Mechanism program (W911NF-14-1-0364).

Bibliographical References

Abu-Jbara, A. and Radev, D. (2011). Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 500–509. Association for Computational Linguistics.

Cai, S. and Knight, K. (2013). Smatch: an Evaluation Metric for Semantic Feature Structures. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 748–752.

Councill, I. G., Giles, C. L., and Kan, M.-Y. (2008). ParsCit: an Open-source CRF Reference String Parsing Package. In *Proceedings of Language Resources and Evaluation Conference*.

Dunne, C., Shneiderman, B., Dorr, B., and Klavans, J. (2010). iopener workbench: tools for rapid understanding of scientific literature. In *Proc. 27th Annual Human-Computer Interaction Lab Symposium*.

Dunne, C., Shneiderman, B., Gove, R., Klavans, J., and Dorr, B. (2012). Rapid understanding of scientific paper collections: integrating statistics, text analytics, and visualization. *JASIST: Journal of the American Society for Information Science and Technology*.

Gupta, S. and Manning, C. D. (2011). Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers. In *Proceedings of the International Joint Conference on Natural Language Processing*.

Kiela, D., Guo, Y., Stenius, U., and Korhonen, A. (2015). Unsupervised discovery of information structure in biomedical documents. *Bioinformatics*, 31(7):1084–1092.

Liu, Y., Bai, K., Mitra, P., and Giles, C. L. (2007). Tableseer: automatic table metadata extraction and searching in digital libraries. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 91–100. ACM.

Marcu, D., Wang, W., Echihabi, A., and Knight, K. (2006). SPMT: Statistical machine translation with syntactified target language phrases. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 44–52, Sydney, Australia, July. Association for Computational Linguistics.

Qazvinian, V., Radev, D. R., Mohammad, S. M., Dorr, B., Zajic, D., Whidby, M., and Moon, T. (2013). Generating extractive summaries of scientific paradigms. *Journal of Artificial Intelligence Research*, 46:165–201.

Radev, D. R., Muthukrishnan, P., and Qazvinian, V. (2009). The ACL anthology network corpus. In *Proceedings, ACL Workshop on Natural Language Processing and Information Retrieval for Digital Libraries*, Singapore.

Schäfer, U., Kiefer, B., Spurk, C., Steffen, J., and Wang, R. (2011). The ACL anthology searchbench. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations*, pages 7–13. Association for Computational Linguistics.

Shahaf, D., Guestrin, C., and Horvitz, E. (2012). Metro maps of science. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1122–1130. ACM.

Teufel, S. and Kan, M.-Y. (2011). *Robust argumentative zoning for sensemaking in scholarly documents*. Springer.

Tsai, C.-T., Kundu, G., and Roth, D. (2013). Concept-based analysis of scientific literature. In Qi He, et al., editors, *CIKM*, pages 1733–1738. ACM.

[5]http://www.mt-archive.info