Appendix B Extracting Wages from Job Descriptions

The job posting data from LinkUp only contains raw job descriptions and does not contain wage information. I use the following procedure to extract posted wages from text-based job postings. Figure A12 illustrates the procedure graphically.

- **Step 1: Extract text chunks containing a dollar sign followed by a digit (e.g., $12, $9, $52,000) from job descriptions.**

  Since raw job descriptions can be very long and contain a lot of information irrelevant to wage information, I keep only sentences in job postings that contain $ followed by a digit or digits. If a job description contains wage information, posted wages should be in these sentences. Without cutting raw job descriptions into shorter sentences, Step 2 can be very time-consuming.

- **Step 2: Use a finetuned question-answering transformer to extract text segments containing wage information from text chunks.**

  After obtaining the sentences that may contain wage information, I use a question-answering transformer to extract phrases that contain posted wages from those sentences. Transformers are a type of neural network architecture that has gained widespread use in NLP tasks such as language modeling, translation, and question-answering. They were first introduced in 2017 by Vaswani et al. (2017).

  A question-answering transformer requires two inputs to extract an answer: a question and a context. The transformer will extract the answer to the question from the context and produce a confidence score of the answer. The score ranges from 0 to 1. The more confident the transformer is about the extracted answer, the higher the score. If the context does not contain an answer to the question, the transformer will still extract an "answer", but the score will usually be close to 0.

  I finetuned a pre-trained transformer to achieve better performance in extracting wage information from job descriptions. The pre-trained model used for finetuning is `deberta-v3-large-squad2`. This model is trained using a large set of English Wikipedia articles and has learned general-purpose representations of language that can be finetuned for the downstream task, wage extraction, with relatively little labeled job postings data. The pre-trained transformer learns domain-specific language patterns in job postings during finetuning.

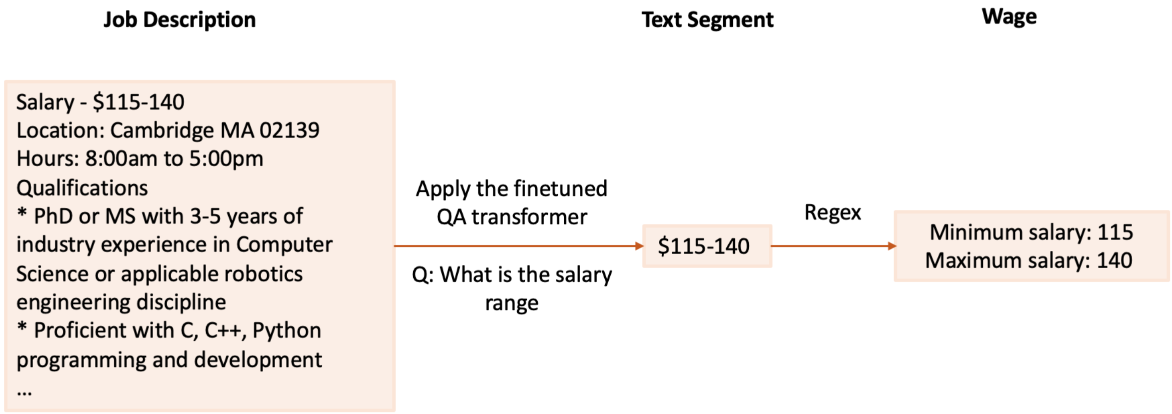
  I randomly drew about 18000 job postings and constructed a labeled dataset with job descriptions and correct wage information for finetuning. The question input of the transformer is "What is the salary range". After finetuning, the accuracy of the transformer improved remarkably. Table A7 shows the evaluation metrics of the original and the finetuned models. The F1 score increased from 66 to 94, and exact matches increased from 54% to 88%.

  In rare cases, the text segment containing the complete wage range is too long to be extracted by the transformer. For example, if the text segment containing the full range is "*Range minimum: $18.00 /hr + bonus * Range maximum: $31.00 /hr + bonus", the answer generated by the transformer to the question will be "Range maximum: $31.00 /hr + bonus". For these cases, I change the question from "What is the salary range" to "What is the maximum salary" and "What is the minimum salary" and apply the finetuned transformer to extract maximum and minimum salaries separately.

- **Step 3: Use a regular expression to extract wage numbers from text segments containing wage information.**

  After getting text segments containing wages, I use a regular expression to extract all numbers following a dollar sign in text segments. I code the smallest number as the minimum salary and the largest as the maximum salary.

Figure A12: Procedure of Wage Extraction from Job Descriptions


Job Description
Text Segment
Wage
Salary - $115-140
Location: Cambridge MA 02139
Hours: 8:00am to 5:00pm
Qualifications
* PhD or MS with 3-5 years of industry experience in Computer Science or applicable robotics engineering discipline
* Proficient with C, C++, Python programming and development
...
Apply the finetuned QA transformer
Q: What is the salary range
$115-140
Regex
Minimum salary: 115
Maximum salary: 140

Note: This figure illustrates the procedure of extracting wages from text-based job descriptions using an excerpt of a job posting.

Table A7: Evaluation Metrics of the Finetuned and the Original Transformers

Metric	Finetuned	Original
% Exact Match	88.05	54.08
F1 Score	93.76	66.39
Sample Size	2694	2694

Note: This table compares the performance of the fintuned and the original transformers. For each question+answer pair, if the characters of the model's prediction exactly match the characters of (one of) the True Answer(s), Exact match = 1; otherwise, Exact match = 0. The F1 score is the harmonic mean of the precision and recall.

References

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.