|  |  |
| --- | --- |
| Title | Deep Learning for Entity Matching: A Design Space Exploration |
| Author | Sidharth Mudgal,Han Li,Theodoros Rekatsinas,AnHai Doan,Youngchoon Park  ,Ganesh Krishnan,Rohit Deep,Esteban Arcaute,Vijay Raghavendra |
| URL | http://pages.cs.wisc.edu/~anhai/papers1/deepmatcher-sigmod18.pdf |
| Content | 3 kinds of EM problem: structured data instances, textual instances, dirty instances  4 kinds of DL methods: SIF, RNN, Attention, Hybrid  The results show that DL does not outperform current solutions on structured EM, but it can significantly outperform them on textual and dirty EM.  **Experiment:**  11 EM tasks for structured instances, 6 EM tasks for dirty instances, 6 tasks for dirty instances with the number of labeled instances ranging from 450 to 250k  It compare the four DL solutions( SIF, RNN, Attention, Hybrid) with Magellan( a state-of-the-art open-source learning-based EM solution  **Result:**  Our results show that DL solutions are competitive with Magel-lan on structured instances (87.9% vs 88.8% average F 1 ), but require far longer training time (5.4h vs 1.5m on average). Thus, it is not clear to what extent DL can help structured EM (compared to just using today learning-based EM solutions)  DL significantly outperforms Magellan on textual EM, improving ac-  curacy by 3.0-22.0% F1 . Our results also show that DL significantly  outperforms Magellan on dirty EM, improving accuracy by 6.2-32.6% F1 . Thus, DL proves highly promising for textual and dirty  EM, as it provides new automatic solutions that significantly out-  perform current best automatic solutions.    **Contributions:**   * We provide a categorization of DL solutions for numerous   matching tasks, and define a design space for these solutions,  as embodied by four DL solutions SIF, RNN, Attention, and  Hybrid.   * We provide a categorization of EM problems into structured   EM, textual EM, and dirty EM   * We provide an extensive empirical evaluation that shows that   DL does not outperform current EM solutions on structured  EM, but it can significantly outperform them on textual and  dirty EM.   * We provide an analysis of DL’s performance and a discussion   of opportunities for future research.  **Design of DL solution:**    **The main takeaways**   * When a limited amount of training data is available,   models that use soft alignment (see Section 3.3) during attribute summarization should be preferred as they yield up to 23% higherF1 over simpler DL models (see Section 5.4.2).   * When a lot of training data is available, the accuracy difference between complex and simpler DL models is smaller.   Thus, one can use simpler DL models that are faster to train (see Section 5.4).  **Understanding What DL Learns**  details:  Sidharth Mudgal et al. 2018. Deep Learning For Entity Matching: A Design Space Exploration. Technical Report. <http://pages.cs.wisc.edu/~anhai/papers/deepmatcher-tr.pdf>.  **Opportunities**   * Our DL results suggest that DL can be promising for many problems in the broader field of data integration (DI) * An exciting future direction is to design simple rule-based optimizers that would analyze the EM task at hand and automate the deployment of such DL models. * A promising research direction is to explore mechanisms for introducing domain-specific knowledge to DL models. |
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|