

Query Spelling Correction Using Multi-task Learning

Xu Sun
Dept. of Statistical Science
Cornell University
Ithaca, NY 14853
xusun@cornell.edu

Anshumali Shrivastava
Dept. of Computer Science
Cornell University
Ithaca, NY 14853
anshu@cs.cornell.edu

Ping Li
Dept. of Statistical Science
Cornell University
Ithaca, NY 14853
pingli@cornell.edu

ABSTRACT

This paper explores the use of online multi-task learning for search query spelling correction, by effectively transferring information from different and biased training datasets for improving spelling correction across datasets. Experiments were conducted on three query spelling correction datasets, including the well-known TREC benchmark data. Our experimental results demonstrate that the proposed method considerably outperforms existing baseline systems in terms of accuracy. Importantly, the proposed method is about one-order of magnitude faster than baseline systems in terms of training speed. In contrast to existing methods which typically require more than (e.g.,) 50 training passes, our algorithm can very closely approach the empirical optimum in around five passes.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Query Formulation*

General Terms

Algorithms, Experimentation, Performance

Keywords

Query Spelling Correction, Multi-task Learning

1. INTRODUCTION

Since query spelling correction is important and received much attention [1, 5, 4, 3], a variety of datasets were developed (from different domains) for this task. As discussed in [2], queries in one specific dataset can be very biased from another dataset, and combining such biased datasets is in general difficult, because (i) their error-patterns can be high biased; (ii) their distributions of misspelled queries can be quite different; (iii) their domains can be quite different (hence, a domain adaptation problem exists). Naturally, the question is: how to effectively integrate such highly biased datasets for improving query spelling correction?

In our paper, we propose a multi-task learning framework for query spelling correction. Our method is able to adaptively integrate highly biased training datasets via automatically learning the task-relationships (data-similarities). As

a result, our method is capable of “softly” merging biased datasets for learning a unified speller. In our approach, we first generate a large amount of spelling correction candidates via a candidate generation module, and then use multi-task learning for accurately re-ranking top candidates.

2. METHODOLOGY

The first step is to generate correction candidates. Following Ganjisaffar *et al.* [2], we implemented candidate generators in our system that generate about 2,000 candidates for each query. Then, we perform naive ranking of the candidates via the Web n-gram service provided by Microsoft.

2.1 Re-Ranking

On top of the naive ranker, a re-ranker based on the *Maximum Entropy* model can significantly improve the results from the naive ranker. Assuming a feature function that maps a pair of observation sequence \mathbf{x} and a classification label y to a feature vector \mathbf{f} , the probability is modeled as

$$P(y|\mathbf{x}, \mathbf{w}) = \frac{\exp[\mathbf{w}^\top \mathbf{f}(y, \mathbf{x})]}{\sum_{y'} \exp[\mathbf{w}^\top \mathbf{f}(y', \mathbf{x})]}, \quad (1)$$

where \mathbf{w} is a weight vector. Given a training set, (\mathbf{x}_i, y_i) for $i = 1 \dots n$, weights are gained via maximizing the objective function, $\mathcal{L}(\mathbf{w}) = \sum_{i=1}^n \log P(y_i|\mathbf{x}_i, \mathbf{w}) - R(\mathbf{w})$, where the term $R(\mathbf{w})$ is an L_2 regularizer in this study.

2.2 Features for Re-Ranking

In addition to the features based on transformations between query-candidate pairs, we also design other features from the top candidates. In total, we use 30 feature templates for each query-candidate pair, which include: error model features (e.g., edit distance), candidate language model features, query language model features, surface features capturing differences of the query and the candidate, frequency of the query and the candidate, the rank of the candidate in the naive ranking phase, and so on.

2.3 Multi-Task (Online) Learning (MTL)

Define $\mathcal{N}_q = \{1, \dots, q\}$. Let T be the number of tasks. For each task $t \in \mathcal{N}_T$, there are n data examples $\{(\mathbf{x}_{t,i}, \mathbf{y}_{t,i}) : i \in \mathcal{N}_n\}$ available. \mathbf{D} denotes the $n \times T$ matrix whose t -th column is given by the vector \mathbf{d}_t of data examples, and \mathbf{W} denotes the $f \times T$ matrix whose t -th column is given by the vector \mathbf{w}_t . We learn \mathbf{W} by maximizing the objective:

$$\text{Obj}(\mathbf{W}, \mathbf{D}) = \sum_{t, t' \in \mathcal{N}_T} \left[\alpha_{t, t'} \sum_{i \in \mathcal{N}_n} \ell_{t'}(i, \mathbf{w}_t) \right] - \sum_{t \in \mathcal{N}_T} \frac{\|\mathbf{w}_t\|^2}{2\sigma_t^2},$$

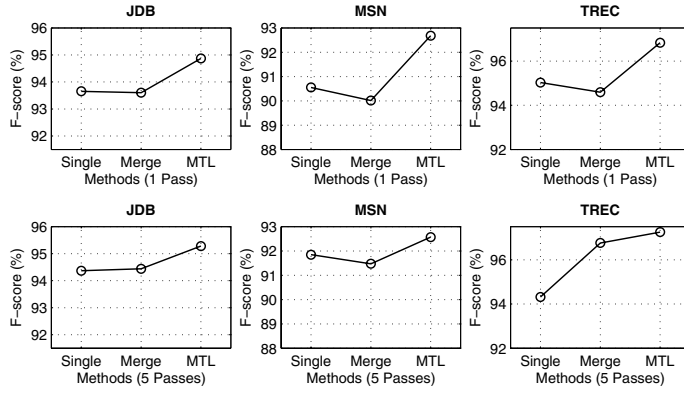


Figure 1: Comparison of F-scores of different methods in 5 passes.

where $\alpha_{t,t'}$ is the *task-similarity*, with $\alpha_{t,t'} = \alpha_{t',t}$. In our study, the similarity is automatically learned from the data.

For online learning, we use the update formula as follows:

$$\mathbf{w}_t^{(k+1)} = \mathbf{w}_t^{(k)} + \eta_t \cdot \mathbf{g}_t, \quad (2)$$

The update term \mathbf{g}_t is derived by weighted sampling over different tasks. η_t is a positive vector-valued step size. The optimal step size asymptotically approaches the (approximate) inverse Hessian matrix.

3. EXPERIMENTS

The TREC dataset is based on the TREC queries (2008 Million Query Track). This dataset contains 5,892 queries, with 5.3% of queries are judged as misspelled. We use two auxiliary datasets for multi-task learning. One is the **JDB** dataset [2] collected from the AOL query sets, with a total of 12,000 samples, and 2,000 of them (16.7%) are misspelled. The second one is collected from **MSN** queries, with about 11% of queries are misspelled. We evenly split the datasets into training and testing data. Baseline methods include the (well-known) stochastic gradient descent method with single-task setting (SGD-Single), and the SGD method with merged setting (SGD-Merge). For the multi-task learning method, its hyper-parameters are tuned in preliminary experiments using cross validation.

As in the 2011 Speller Challenge, we use the expected F-score as the evaluation metric. The performance comparisons on the 1-pass and 5-pass settings are highlighted in Figure 1. In this figure, *Single* represents the *SGD-Single* baseline. Similarly, *Merge* represents *SGD-Merge*. Clearly, the proposed MTL method can achieve substantially higher F-scores than the baselines.

In addition, we check the training process and the empirical convergence speed. The curves are shown in Figure 2. The figure shows the comparisons of MTL with baselines. We find the MTL method has the best empirical convergence speed. The results demonstrate that the proposed method can achieve good accuracy with fast convergence speed.

4. CONCLUSION

We demonstrated that using multi-task learning can reliably integrate highly biased datasets for query spelling correction and may outperform existing baselines in terms of

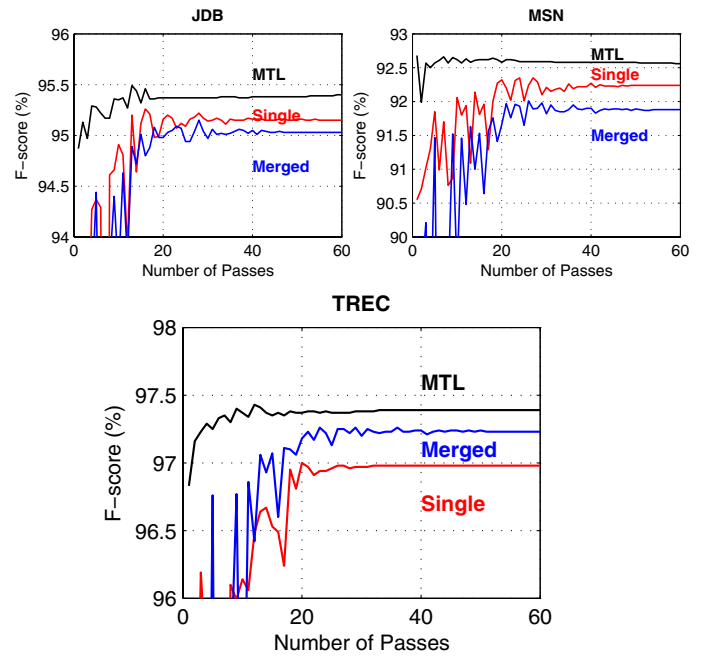


Figure 2: F-score curves of different methods.

accuracy and speed. Since query spelling correction can have very large-scale training data in industrial applications (e.g., query spelling logs which contain millions of entries, see [4]), the proposed method's ability to approach empirical optimum in a few passes will be of critical importance in such large-scale applications.

5. ACKNOWLEDGMENTS

This work is partially supported by NSF (DMS-0808864, SES-1131848), ONR (YIP-N000140910911), and DARPA. Xu Sun was a Postdoctoral Associate supported by ONR and NSF. Anshumali Shrivastava is a Ph.D. student supported by ONR and NSF. The authors thank Yanen Li (Author of [3]) for helpful discussions.

6. REFERENCES

- [1] BRILL, E., AND MOORE, R. C. An improved error model for noisy channel spelling correction. In *ACL'00* (Hong Kong, October 2000), pp. 286–293.
- [2] GANJISAFFAR, Y., ZILIO, A., JAVANMARDI, S., CETINDIL, I., SIKKA, M., KATUMALLA, S. P., KHATIB-ASTANEH, N., LI, C., AND LOPES, C. qSpell: Spelling correction of web search queries using ranking models and iterative correction. In *Spelling Alteration for Web Search Workshop* (July 2011).
- [3] LI, Y., DUAN, H., AND ZHAI, C. Cloudspeller: Spelling correction for search queries by using a unified hidden markov model with web-scale resources. In *Spelling Alteration for Web Search Workshop* (July 2011).
- [4] SUN, X., GAO, J., MICOL, D., AND QUIRK, C. Learning phrase-based spelling error models from clickthrough data. In *ACL'10* (Uppsala, Sweden, July 2010), pp. 266–274.
- [5] TOUTANOVA, K., AND MOORE, R. Pronunciation modeling for improved spelling correction. In *ACL'02* (Philadelphia, USA, 2002), pp. 144–151.