Understanding Scanned Receipts

Eric Melz
300 S Reeves Dr.
Beverly Hills, CA 90212
`eric@emelz.com`

Abstract

Tasking machines with understanding receipts can have important applications such as enabling detailed analytics on purchases, enforcing expense policies, and inferring patterns of purchase behavior on large collections of receipts. In this paper, we focus on the task of Named Entity Linking (NEL) of scanned receipt line items. Specifically, the task entails associating shorthand text from OCRd receipts with a knowledge base (KB) of grocery products. For example, the scanned item "STO BABY SPINACH" should be linked to the catalog item labeled "Simple Truth Organic™Baby Spinach". Experiments that employ a variety of Information Retrieval techniques in combination with statistical phrase detection shows promise for effective understanding of scanned receipt data.

1 Introduction

Tasking machines with understanding receipts can have important applications such as enabling detailed analytics on purchases, enforcing expense policies, and inferring patterns of purchase behavior on large collections of receipts. In this paper, we focus on the task of Named Entity Linking (Hachey et al., 2012) of scanned receipt line items. Specifically, the task entails associating shorthand text from OCRd receipts with a knowledge base (KB) of grocery products. For example, the scanned item "STO BABY SPINACH" should be linked to the catalog item labeled "Simple Truth Organic™Baby Spinach".

2 Related Work

Blah blah blah.

3 Data

Blah blah blah.

4 Methodology

Blah blah blah.

5 Experiments

Blah blah blah.

5.1 Baseline

Blah blah blah

5.2 Wildcards

Blah blah blah

5.3 Mashed Wildcards

Blah blah blah

5.4 Phrases

Blah blah blah

5.5 Fuzzy Phrases

Blah blah blah

6 Results

blah

7 Conclusion

blah

References

Ben Hachey, Will Radford, Joel Notham, Matthew Honnibal, and James R. Curran. 2012. Artificial intelligence, wikipedia and semi-structured resource-sevaluating entity linking with wikipedia. *Artificial Intelligence*, 194(11):130–150.

A Appendices

blah

B Supplemental Material

blah blah blah