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

A literature review reveals virtually no published work in this specific domain. While there is a body of work researching text extraction from scanned receipts (e.g. Huang et al., 2019), the work is primarily focused on Named Entity Recognition (NER) instead of Named Entity Linking (NEL). That is, systems are considered successful if they can identify text items such as store locations, totals, etc, but they are not evaluated with respect to the interpretation of the extracted text.

Although no papers exist on linking scanned entities, there is literature in other areas that appear potentially relevant to the subject task. This includes work on general-purpose techniques for building abbreviation dictionaries, acquisition of medical abbreviations (e.g., "COPD" → "Chronic Obstructive Pulmonary Disorder"), and normalization of social media content (e.g., "ur coooool" → "you are cool"). The follow sections summarize a few papers in these areas.

2.1 Language Independent Acquisition of Abbreviations

(Glass et al., 2017) describe a language-independent technique for acquiring abbreviations and their expansions, by exploiting Wikipedia redirect and disambiguation pages. They begin by motivating the acquisition of abbreviations, noting that the explosion of social media has made the need for abbreviations increasingly important. They also note that a token such as "ACE" could have multiple expansions, including "accumulated cyclone energy" and "American Council on Education" in addition to the word "ace" (as in "Ace of spades"). The authors present related work, noting that most of the previous work for abbreviation detection and expansion extraction has been in the domain of English biomedical text. A common strategy is to identify occurrences where an abbreviation is explicitly paired with its expansion for example through a pattern involving a parenthetical such as `<short form> (<long form>)` or `<long form> (<short form>)`. Other approaches consider the contexts of short form

and long form occurrences, pairing short forms with long forms according to their distributional similarity by measuring the cosine of their context vectors. Another approach uses supervised learning, considering features such as string similarity and other characteristics of the short and long forms. The authors work is based on previous work by (Jacquet et al., 2014) who describe a technique for mining abbreviations by making use of Wikipedia redirection pages. The authors observe that, due to the use of only redirect pages for the gold standard annotation, a shortcoming of the prior work is that each abbreviation only has a single expansion even though multiple different expansions are possible for some of the abbreviations. To remedy this shortcoming, the authors propose mining disambiguation pages in addition to redirect pages to gather multiple possible long-form expansions. The authors mine redirect and disambiguation pages for abbreviations, while applying several rules such as (a) Short forms are restricted to ten characters or less, (b) At least half of the short-form characters must be upper case, and (c) The long-form must be at least twice as long as the short form, with at least two tokens. They generate candidate expansions and then score the expansions. Scoring occurs by computing features for synonym similarity, topic similarity, and surface similarity. Synonym similarity means that one term can be replaced with another while preserving the meaning of the sentence and is assessed using word embeddings using word2vec (Mikolov et al., 2013). Topical relatedness means that two terms occur in the same sorts of documents, and is assessed using Latent Semantic Analysis (Deerwester et al., 1990). Surface similarity is the overlap in the surface forms of the terms by computing the best possible alignment between a short form and a long form. The three similarity scores are combined using a logistic regression model. The authors compare their system with a previous system developed by (Schwartz and Hearst, 2003) that extracts abbreviations using parentheses based patterns. The metric used to compare systems is Area Under the Precision/Recall curve. Without the scoring extensions, the 2 systems are comparable: the Schwartz and Hearst system has an AUC of 0.359 and the Candidate System has an AUC of 0.324. However, by adding the alignment and embedding scoring extensions, the Candidate Systems performance improves to an AUC of 0.480.

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 Future Work

blah

8 Conclusion

blah

References

S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R.A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science 41*, pages 391–407.

Michael R. Glass, Md. Faisal Mahbub Chowdhury, and Alfio Massimiliano Gliozzo. 2017. Language independent acquisition of abbreviations. *CoRR*, abs/1709.08074.

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

Z. Huang, K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, and C. V. Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520.

Guillaume Jacquet, Maud Ehrmann, and Ralf Steinberger. 2014. Clustering of multi-word named entity variants: Multilingual evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Ariel S. Schwartz and Marti A. Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Pacific Symposium on Biocomputing*, pages 451–462.

A Appendices

blah

B Supplemental Material

blah blah blah