Structural Dissection of Business Entity Names

Venkatesh Vinayakarao
Indraprastha Institute of Information Technology
New Delhi
venkateshv@iiitd.ac.in

Srikanta Bedathur
Indraprastha Institute of Information Technology
New Delhi
bedathur@iiitd.ac.in

Amani Kongara
Indraprastha Institute of Information Technology
New Delhi
kongara1240@iiitd.ac.in

ABSTRACT

Business entity (also known as local listing) names pose big challenge for search engines. Business entities like any other entity (people, place or product) appear to be unstructured at the outset and apparantly demonstrate human creativity. Our claim is that there are parts of these names that are recognizable and useful. We inspect 11,537 business entities from Phoenix, USA provided by the Yelp data set[1] towards gaining deep insights into how they are structured and if they can be annotated effectively. We argue about the characteristics of such names in detail including the opportunities and challenges in their annotation. An effort of this kind should lead to recommending a standard for capturing business names. A standard is much necessary in this domain since there are varied data providers and multiple search engines are consumers. Such a standard will alleviate the issues of record linkage, matching and ranking. In this process, we also make available, a manually tagged corpus of entity names for future research.

Categories and Subject Descriptors

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

Keywords

Tagging, Annotation, Normalization, Standardization

1. INTRODUCTION

Local search refers to the search of business in any specific place or their neighbourhoods. Let's say you are on the road and wish to locate the nearest seafood restaurant. In this case, you are doing a local search. With advances in mobile technologies and internet, local search has become indispensable. Business entity names play a significant role in the process of local search.

[1]http://www.yelp.com/dataset_challenge/

At the outset, business entity names seem to have no obvious structure. They are reflections of human creativity and the nature of business. We believe that a structural understanding of names will allow us to recognize, extract and normalize several of its parts. Most commercial search engines such as Google and Bing are known to buy foot-on-the-road entity data. Also, they may have their own proprietary systems to mine the web. Owing to existence of several such data sources (including prominent providers such as yelp and trip-advisor) and the variety in the way they capture the names, each entity gets morphed into several variants and that causes complications in record linkage, matching and ranking. For instance, consider, "Queen Elizabeth's Grammar School for Boys, Barnet, London". This is also known as "QE Boys", "QE Grammar School", "Queen Elizabeth's Grammar School" and "Queen Elizabeth's School". These are just few of the references to this school found on the web. An entity data provider may choose any of such ways to capture this entity, by name.

Towards a successful local search experience, its important to get a deep understanding of entity names. In this paper, we investigate the essential constituents of entity names and attempt to annotate them.

2. RELATED WORK

There has been lot of work done on text document segmentation[1][7] and Query Segmentation [6][5]. Document segmentation typically looks at finding coherent sentences from textual documents typically through statistic, linguistic or probabilistic means. Query segmentation is much closer to our work since the text being segmented are short and have named entities in them. There has also been considerable work done on named entity recognition[4][8] and linguistic processing of short text [11][14]. However, the same techniques that work with queries do not work with entity names. For example, "My Sister Toy Shop" should lead to "My Sister" as the core name and "Toy Shop" as the category element. Most of the POS taggers will identify "My" as a Preposition which is not at all the case in this context. "A Stitch in Time" makes custom fit clothes. This name should not be split. Entire text, together, identifies the business.

Takahashi et al. [13], analyze the truthfulness of modifiers such as "authentic", "impressive" in entity names. Oates et al.,[12] look for same entities in free text using LSA based approach. There has been several attempts at normalizing text such as gene name normalization[3], person name nor-

malization[10] and location normalization [9]. Business entity name normalization depends on these but is much more complicated since names could be a combination of person name, location, domain, etc.

Bouquet et al.,[2] proposed ENS (Entity Name System) to refer names uniformly. Their approach is to hash the names to uniform identifier. For example, "Paolo Bouquet" is referred to as "ok20070630118579125 2056" and is geared towards generation of such an identifier. One of their requirements is to maintain a large repository of global entity names. We take a different perspective of normalizing names. Our work could serve as a pre-processing step to this kind of normalization. While unique representation for a name happens to be the objective of Bouquet et al's approach, our approach works towards improving IR by extracting key elements of a name.

Our work is to extract, re-organize and normalize the name parts. This does not always result in the same unique representation. Yet, this approach does generate semantically richer names. Humans can identify that "Nimbus American Bistro N' Brewery" refers to an entity whose name might be "Nimbus" or "Nimbus American" while belonging to "Bistro" and "Brewery" categories. We believe this can be learned and such names can be machine translated.

3. BUSINESS ENTITY NAME

A typical entity name comprises of a core name along with various segments such as business categories and location. Since naming of an entity demands human creativity, there are several forms that an entity name can take. Essentially, we observed the entity name to contain these high level items:

1. Location: In "75th Thai Taste Restaurant", "75th" is part of the address and is followed by the core name "Thai Taste" whereas "Restaurant" is the category. The Location field may occur in the beginning or the later part along with the core name of an entity name.

2. CoreName: The name, Scottsdale Medical Imaging LTD, has Scottsdale as the core name of the entity that distinguishes this entity from rest of the entities in the "Medical Imaging" domain. In the example "A Stitch in Time", the whole string is a core name. CoreName plays a huge role in identification of an entity and thus is a crucial component of entity name. In the manually annotated training set, we observed that 66% of these CoreNames where either a God or Person.

3. Category: In "Every Kid's Dentist & Orthodontics", "Every Kid's" is the core name of the entity which distinguishes it among the other entities of the categories, Dentists and the Orthodontics. In "Autohaus Service & Performance Center", "Autohaus" is the core name, whereas "Service" & "Performance Centre" is a category. Most of the entity names occur with the category to which it is associated.

4. Anchor: In "KC & Co.", "KC" is the core name and "Co." is the anchor. Similarly in "Phoenix Paediatrics Ltd", "Phoenix" is the core name, "Paediatrics" is its domain and "Ltd" is its anchor word. HongKong's Company Naming Guideline [2] says that these anchors ("Limited", "Company", "Company Limited", etc) will be ignored while checking if the companies are same.

5. Others: "R Bar at the Camelback Inn" contains the core name "R Bar", the connector "at" and the address "The Camelback Inn". We bucket such items that do not fall into any other recognizable part as "others".

We wondered if there exist more ways of categorizing the parts of entity names. These parts came out of a 5 member survey on 100 names each. Each of them identified the parts in the similar way. As we see an entity name can be of any combination of these parts. Hence, their annotation is an interesting and challenging task.

4. NAMES AS HIDDEN MARKOV MODEL

We visualize the words in entity name as a sequence of observed states. Behind each word, we visualize its category as its hidden state. We prepared the manually annotated training data with 500 randomly chosen entity names. A model thus generated was used with Viterbi decoding algorithm to predict the best possible hidden states, thereby, giving us the annotations.

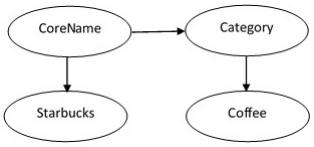

Figure 1: Visualization of our annotator as HMM

The HMM is described with a state space S with initial probability pi, x_i observed states and y_i hidden states. In our case there are n observed and hidden states where n is the number of tokens in the entity name. If α denotes the transition probability between states of S and k is the last state, the most likely state sequence is computed by viterbi method as

$$V_{t,k} = \mathrm{P}\big(y_t \mid k\big) \cdot \max_{x \in S}(a_{x,k} \cdot V_{t-1,x})$$

The transition probabilities learnt from manual annotations were very interesting. For instance, in all our examples, only anchor followed an anchor if at all there were multiple anchor terms. Hence the probability is 1. Category has even distribution for all transitions while corename shows a bias towards leading to categories. We believe these numbers are dependent on the culture and reflects what is considered as "most common". All our data belongs to businesses in Phoenix, USA.

Beyond an extent, precision could not be increased using training data. In the case of core names, the lack of precision attributed majorly to their abbreviations, acronyms, multilanguage nature and varied positioning in the word sequence. This model seems to work fine with category much better

[2]http://www.cr.gov.hk/en/publications/docs/name-e.pdf

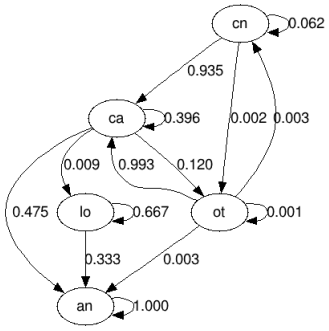

Figure 2: Transition probabilities between annotations from training data

Part	Precision
CoreName	71%
Category	82%
Location	72%

Table 1: Accuracy of HMM based annotations

than the rest. Categories, by nature are dictionary terms (such as school, university, hospital, mall, theatre, etc) and they repeat. Identifying category first and moving on to use this knowledge to fix the rest of the parts gave better results. Full list of country-specific anchors (such as pvt, ltd, co, inc, gmbh, etc) are available and hence could be hard-wired into our system. Hence all the evaluation results are given only for CoreName, Category and Location. Further, note that names can have multiple tokens as core name, category or location. For evaluation, we consider a result as precise only when all the manually judged corename tokens are identified as corename by this system. Same applies to location and category. For example, if judges have marked "Phoenix/CN Children's/CA Hospital/CA", we need the system to mark both Children's and Hospital tokens as category inorder to be counted as correct.

5. BOOSTING THE HMM RESULTS

Along with HMM based name part detection, we observed following opportunities to boost precision:

1. URL Breaking: URLs give some indication of which part of the entity name is core name. For example, "Clearly Professional Window Cleaning" is www. cpwindowcleaning.com. Looking at several names, we can say that "window cleaning" is a popular category and "Clearly Professional" or "CP" is what uniquely identifies them.

2. Address helps us resolve the location part of the name (if any) and also to detect LIE (location in entity). For example,"Vijayawada Public School, Vijayawada" is both an LIE as well as has location at the end, in its name.

3. "Arizona Auto Care" could be searched as "Arizona car care", "Arizona car service", "Arizona truck repair" and so on. There should be a way to extract these synonyms. A look at several companies in the auto care business allows us to capture a set of synonyms for "auto care".

4. Abbreviations and Acronyms: BBQ, bnb, McD, Caltech, zoo vs zoological park. These can be found not only in core name but any part of the name.

5. Name Breaking: PetSmart could be searched as "Pet Smart". We keep the investigation of related work on decompounding as future work. For now, we rely purely on capitalization.

6. Multi-Language nature - Its common to see chinese restaurants in India, Indian Restaurants in US and French restaurants everywhere!. Its fair to expect that the names will carry multiple languages. "Tortas El Guero", Gorditas El Tio are examples.

7. Normalization: Each constituent needs to be normalized. For example, Mumbai and Bombay refer to same cities. "Satyam Computers" became "Tech Mahindra". Candidates for normalization also include words such as 'N, and, & etc. For now, we have used a manual set of translations to normalize common words/phrases such as "and" for "'N", "BBQ" for "Barbeque" and so on. As future work, we leave creation of a synset for each part of the name and associate it to the normalized business name.

For now, we used existing data from wikipedia and other internet sources to compile a supporting list of data such as location names, anchors and synonyms. This data was used to boost the hmm results. We observed a gain in precision as given in the table below.

Part	Precision
CoreName	77%
Category	84%
Location	83%

Table 2: Accuracy of Boosted annotations

Some examples of the normalized text looks as follows:

In the above table, annotations expand as follows: CN:CoreName, CA:Category, LO:location, AN:Anchor, OT:Others.

Name	Annotated Name
A Stitch in Time	A/CN Stitch/CN in/CN Time/CN
Scottsdale Medical Imaging LTD	Scottsdale/CN Medical/CA Imaging/CA LTD/AN
California Institute of Technology	California/CN Institute/CA of/CA Technology/CA

Table 3: Sample Annotations

6. CONCLUSION AND FUTURE WORK

[16] suggests with evidence that the existing name extraction approaches are not good enough. Our work should help such efforts. Moreover, the heuristics we looked at for boosting the results can be more elegantly modeled using markov logic networks. The training data that we used has 500 names with tags that are judged to be correct which is released for further research and we hope to build a bigger corpus.

Firstly and most importantly, it will be interesting to see the Precision, Recall and F-Scores for Learning to rank data sets when the annotations are available as additional features. Secondly, language issues in name need to be considered. We worked only with English names and names that have non-dictionary words as category were left out. Thirdly, we stopped with 2 levels of category description (domain, category) considering that we had only 11k items in our dataset. In real world datasets, the category ontology extraction will be a bigger challenge. There is also scope for improvement with spellers, word decompounders and location repositories. Multiple knowledge sources and non-local data can be used to improve entity identification[15].

From this work, we believe a standard representation of entity names can be achieved and we showed one way of doing it with reasonable accuracy on the Yelp dataset.

7. REFERENCES

[1] D. Beeferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. *Mach. Learn.*, 34(1-3):177–210, Feb. 1999.

[2] P. Bouquet, H. Stoermer, C. Niederee, and A. Maña. Entity name system: The back-bone of an open and scalable web of data. In *Proceedings of the 2008 IEEE International Conference on Semantic Computing*, ICSC '08, pages 554–561, Washington, DC, USA, 2008. IEEE Computer Society.

[3] H.-r. Fang, K. Murphy, Y. Jin, J. S. Kim, and P. S. White. Human gene name normalization using text matching with automatically extracted synonym dictionaries. In *Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*, BioNLP '06, pages 41–48, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[4] J. Guo, G. Xu, X. Cheng, and H. Li. Named entity recognition in query. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 267–274, New York, NY, USA, 2009. ACM.

[5] M. Hagen, M. Potthast, B. Stein, and C. Braeutigam. The power of naive query segmentation. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 797–798, New York, NY, USA, 2010. ACM.

[6] M. Hagen, M. Potthast, B. Stein, and C. Bräutigam. Query segmentation revisited. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 97–106, New York, NY, USA, 2011. ACM.

[7] X. Huang, F. Peng, D. Schuurmans, N. Cercone, and S. E. Robertson. Applying machine learning to text segmentation for information retrieval. *Inf. Retr.*, 6(3-4):333–362, Sept. 2003.

[8] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee. Twiner: Named entity recognition in targeted twitter stream. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 721–730, New York, NY, USA, 2012. ACM.

[9] H. Li, R. K. Srihari, C. Niu, and W. Li. Location normalization for information extraction. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[10] W. Magdy, K. Darwish, O. Emam, and H. Hassan. Arabic cross-document person name normalization. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, Semitic '07, pages 25–32, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[11] E. L. Murnane, B. Haslhofer, and C. Lagoze. Reslve: Leveraging user interest to improve entity disambiguation on short text. In *Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 1275–1284, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[12] T. Oates, V. Bhat, and V. Shanbhag. Using latent semantic analysis to find different names for the same entity in free text. In *Proceedings of the 4th International Workshop on Web Information and Data Management*, WIDM '02, pages 31–35, New York, NY, USA, 2002. ACM.

[13] R. Takahashi, S. Oyama, H. Ohshima, and K. Tanaka. Evaluating truthfulness of modifiers attached to web entity names. In *Proceedings of the 11th International Conference on Web-age Information Management*, WAIM'10, pages 429–440, Berlin, Heidelberg, 2010. Springer-Verlag.

[14] I. Taksa, S. Zelikovitz, and A. Spink. Using web search logs to identify query classification terms. In *Proceedings of the International Conference on Information Technology*, ITNG '07, pages 469–474, Washington, DC, USA, 2007. IEEE Computer Society.

[15] M. Vilain, J. Huggins, and B. Wellner. Sources of performance in crf transfer training: a business name-tagging case study. In *Proceedings of the International Conference RANLP-2009*, pages 465–470, Borovets, Bulgaria, September 2009. Association for Computational Linguistics.

[16] M. Vilain, J. Su, and S. Lubar. Entity extraction is a boring solved problem: Or is it? In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, NAACL-Short '07, pages 181–184, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.