The role of domain in learning about referents

Cassandra L. Jacobs
Stitch Fix Inc.
San Francisco, CA, USA
cjacobs@stitchfix.com

Abstract

A number of natural language processing tasks such as knowledge representation, coreference resolution, or other information retrieval tasks depend on learning good representations of real-world entities. Here we demonstrate that the language associated with different entities is highly sensitive to the domain. By analyzing the entities associated with requests, reviews, and descriptions, we can learn entity representations that are stable across different textual domains, but that the entities that are relevant to a request are not necessarily easily translated into entities that are similar based on how they are described in reviews.

1 Introduction

In an information retrieval context, we often want to show relevant documents or links, such as recipes when an individual types "kale salad" into a research engine. Often, however, the domain in which a model was trained (e.g. sets of recipes) is different from the way users search for items, as very few queries are similar to texts. This is a very common problem in information retrieval, and is especially difficult when handling vague queries.

In this paper, we attempt to quantify how reviews, descriptions, and requests associated with the same real-world entities like SPINACH can provide different, but similar, semantic representations. We illustrate this by showing that an entity like SPINACH will be similar to the same types of entities across domains (e.g. KALE or BROCCOLI) but also that the way we talk about SPINACH differs across different text sources. In this work we examine three different document types: written requests, descriptions of the item, and reviews. We discuss below each of these domains, which provide distinct views into how speakers talk about real-world entities.

Written requests can be conceptualized as long queries. Requests have a number of different properties from ordinary written language (Pilegaard, 1997), and in some domains vary from the specific to the general (Blum-Kulka, 1987). Furthermore, in the case of free-form requests, a person may be offering an indefinite or very large number of objects to be retrieved. Descriptions from encyclopedic entries like Wikipedia, by contrast, are useful for building entity ontologies despite being fairly unstructured. Descriptions are potentially interesting because they often capture the redundant, structural properties of a referent that are less commonly mentioned in other domains (Dale and Reiter, 1995). Reviews constitute relatively structured, referent-linked content (Jo and Oh, 2011) but may fail to mention redundant properties of an entity like a prototypical color. Given these biases, it would be easy to learn, for example, that spinach is "gross" but less straightforward to learn that it is typically green (Willits et al., 2015).

These domain differences are especially salient in an information retrieval context. It is well-understood that queries and the content intended to match the queries differ substantially in the vocabulary used (Lafferty and Zhai, 2001). An informal investigation into this by the authors found that requests can be distinguished between reviews and descriptions in approximately 90% of all cases. In light of these domain differences, it is important to specify how what we might learn about entities can change in different discourse contexts.

2 Dataset

Our entities corpus contains (1) reviews made about different categories of fashion retail items (*reviews*), (2) small notes that describe up to five known items (*descriptions*), and (3) up to five known items that may be an attempt to fulfill a request (*requests*).

To extract meaningful relationships between text and entities, we create bag-of-words and bag-of-entities representations for each document. The entities associated with reviews are treated as one-hot encodings as they are reviews of only a single entity. In contrast to reviews, we assume that any of the five entities associated with those documents may be relevant in descriptions because the author is not obligated to mention any or all of the entities; likewise, we do not know whether any item actually fulfilled a request. For simplicity, we assume that each of the 5 things linked to descriptions and requests are relevant. For these more complex documents we can think of the entity sets associated with each document as bags of entities.

3 Representation learning

We train three separate entity embeddings models on each of these datasets and use the entity representations that we obtain from this process to compare entities across discourse contexts. We assume that entities that are similar in how they are described, or whether they are associated with similar requests, or whose reviews are similar, should have similar representations to each other.

To learn "entity" embeddings, we employ the positive pointwise mutual information approach of (Levy and Goldberg, 2014), but add "bag-of-entities" vectors as described above as features in addition to bag-of-words features. Learned entity embeddings therefore reflect how commonly an entity like SPINACH is associated with a certain word (e.g. 'greens'). By training in this way we can obtain a lower-dimensional, dense representation of each entity that is a function of the words used in documents associated with it. These representations put all entities into a common vector space and make it easy to compare entities by examining how close two entities like SPINACH and KALE are across different linguistic contexts using metrics like cosine distance.

4 Evaluation

One possible means of evaluation is by simulating an information retrieval (IR) task. In a typical IR task, a simple text query (e.g. a single word like *greens*) is generated. This query might return potential ingredients like SPINACH, KALE, or BROCCOLI at the top. If words are used similarly across requests, descriptions, and reviews, then the top matches for *greens* will be similar regardless of how the model was trained (Furnas et al., 1988).

We assess the similarity between two entities by measuring the size of the intersection of the top 10 most similar entities to each entity.

We have two evaluation phases. First, we can take SPINACH in a review context and compare it to SPINACH in a request context and see how similar the top 10 most similar entities are to them (i.e. SAME versus DIFFERENT textual domains). If the entities are highly similar, then we expect to see e.g. KALE and BROCCOLI rank as very similar to SPINACH across domains.

Second, we can compare two different entities to each other within the same text domain. For example, we can see how similar SPINACH is to KALE in a review context. We should see that the 10 closest matches to SPINACH differ from the 10 closest matches to KALE. Because we believe that the text domain affects the features we can learn about entities, we should also expect that the similarity between and two entities like KALE and SPINACH will vary substantially across the different domains.

To see whether discourse context affected whether two entities were likely to be considered similar or not, we compared the average overlap between pairs of entities across six categories, analyzing over 56,000 pairwise observations. We considered two factors – either the referent was the same across two different discourse contexts (e.g. SPINACH_REQUEST and SPINACH_REVIEW) or the same discourse contexts but different entities (e.g. SPINACH_REQUEST and KALE_REQUEST).

Pair type	Same	Different
Review vs. request	1.75	0.25
Request vs. description	2.69	1.95
Review vs. description	3.78	0.65

Table 1: Average number of best matching results in the top 10 for two different linguistic contexts. Same-entity pairs are more similar than different-entity pairs. At the same time, there are substantial differences in the number of matches across discourse contexts.

The results of this analysis demonstrate that we can identify similar entities across domains – spinach is more similar to spinach than it is to kale.

At the same time, it is clear that the entity representations we learn differ substantially from one domain to another. Reviews are not very similar to requests, but they are very similar to descriptions. Requests are similarly somewhat similar to descriptions.

5 Conclusion

We have demonstrated here that the domain from which we learn entity representations has a large impact on the quality of representations we learn. In some text domains, for example, the most similar entities to SPINACH may complement it (e.g. DRESSING), rather than be physically similar (e.g. KALE) (Gentner and Brem, 1999).

As part of this work, we proposed one possible method of evaluating the degree to which unstructured texts like reviews, descriptions, and requests are similar by comparing the overlap between their best matches. Our results suggest that there is greater need to identify the role that different corpora have on our ability to extract knowledge about the real world for natural language processing tasks.

References

Shoshana Blum-Kulka. 1987. Indirectness and politeness in requests: Same or different? *Journal of Pragmatics* 11(2):131–146.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science* 19(2):233–263.

George W Furnas, Scott Deerwester, Susan T Dumais, Thomas K Landauer, Richard A Harshman, Lynn A Streeter, and Karen E Lochbaum. 1988. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pages 465–480.

Dedre Gentner and Sarah K Brem. 1999. Is snow really like a shovel? distinguishing similarity from thematic relatedness. In *Proceedings of the Twenty-First Annual Meeting of the Cognitive Science Society*. Lawrence Erlbaum Associates Mahwah, NJ, pages 179–184.

Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. ACM, pages 815–824.

John Lafferty and Chengxiang Zhai. 2001. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pages 111–119.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*. pages 2177–2185.

Morten Pilegaard. 1997. Politeness in written business discourse: A textlinguistic perspective on requests. *Journal of Pragmatics* 28(2):223–244.

Jon A Willits, Michael S Amato, and Maryellen C MacDonald. 2015. Language knowledge and event knowledge in language use. *Cognitive psychology* 78:1–27.