Mickey Mouse is not a Phrase: Improving Relevance in E-Commerce with Multiword Expressions

Prathyusha Senthil Kumar, Vamsi Salaka, Tracy Holloway King, and **Brian Johnson**
Search Science
eBay, Inc.
San Jose, CA, USA
{ prathykumar, vsalaka, tracyking, bjohnson } @ebay.com

Abstract

We describe a method for detecting phrases in e-commerce queries. The key insight is that previous buyer purchasing behavior as well as the general distribution of phrases in item titles must be used to select phrases. Many multiword expression (mwe) phrases which might be useful in other situations are not suitable for buyer query phrases because relevant items, as measured by purchases, do not contain these terms as phrases.

1 Phrase MWE in e-Commerce Search

Processing buyers' queries is key for successful e-commerce. As with web search queries, e-commerce queries are shorter and have different syntactic patterns than standard written language. For a given query, the system must provide sufficient recall (i.e. return all items relevant to the buyers' query, regardless of the tokens used) and sufficient precision (i.e. exclude items which are token matches but not relevant for the query). This paper looks at how identifying phrases in buyer queries can help with recall and precision in e-commerce at eBay. We focus primarily on precision, which is the harder problem to solve.

Phrases are a sub-type of mwe: one where the tokens of the mwe appear strictly adjacent to one another and in a specified order ((Sag et al., 2002)'s words with spaces).

The eBay product search engine takes buyer queries and retrieves items relevant to the buyer's purchasing intent. The items are listed in categories (e.g. women's dresses) and each item has a title provided by the seller. The buyer can choose to sort the items by most relevant (e.g. similar to web search ranking) or deterministically (e.g. price low to high). There are versions of the e-commerce site for different countries such as US, UK, Germany, France, Poland, etc. and so the query processing is language-specific according to site. Here we report on incorporating phrases into English for the US and German for Germany.

2 Controlling Retrieval via Query Phrases

The query processing system has three core capabilities[1] which expand tokens in the buyer's query into other forms. Both single and multiple tokens can be expanded. Token-to-token expansions (Jammalamadaka and Salaka, 2012) include acronyms, abbreviations, inflectional variants (e.g. *hats* to *hat*), and space synonyms (e.g. *ray ban* to *rayban*). Category expansions expand tokens to all items in a given category (e.g. *womens shoes* retrieves all items in the Womens' Shoes category). Finally, attribute expansions map tokens to structured data (e.g. *red* retrieves any item with Color=Reds in its structured data). These expansions are used to increase the number of relevant items brought back for a specific buyer query.

Precision issues occur when a buyer's query returns an item that is a spurious match. For example, the query *diamond ring size 10* matches all the tokens in the title "10 kt gold, size 7 diamond ring" even though it is not a size 10 ring.

Recall issues occur when relevant items are not returned for a buyer's query. The core capabilities of token-to-token mappings, category mappings, and attribute mapping largely address this. However, some query tokens are not covered by these capabilities. For example, the query *used cars for sale* contains the tokens *for sale* which rarely occur in e-commerce item titles.

[1] Here we ignore tokenization, although the quality of the tokenizer affects the quality of all remaining components (Manning et al., 2008).

Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014), pages 62–66,
Gothenburg, Sweden, 26-27 April 2014.

2.1 Hypothesis: Phrasing within Queries

To address these precision and recall issues, we provide special treatment for phrases in queries. To address the precision issue where spurious items are returned, we require certain token sequences to be treated as phrases. For example, *size 10* will be phrased and hence only match items whose titles have those tokens in that order. To address the recall issue, we identify queries which contain phrases that can be dropped. For example, in the query *used cars for sale* the tokens *for sale* can be dropped; similarly for German *kaufen* (buy) in the query *waschtrockner kaufen* (washer-dryer buy). For the remainder of the paper we will use the terminology:

- REQUIRED PHRASES: Token sequences required to be phrases when used in queries (e.g. *apple tv*)
- DROPPED PHRASES: Phrases which allow sub-phrase deletion (e.g. *used cars for sale*)

The required-phrases approach must be high confidence since it will block items from being returned for the buyer's query.

We first mined candidate phrases for required phrases and for dropped phrases in queries. From this large set of candidates, we then used past buyer behavior to determine whether the candidate was viable for application to queries (see (Ramisch et al., 2008) on mwe candidate evaluation in general). As we will see, many phrases which seem to be intuitively well-formed mwe cannot be used as e-commerce query phrases because they would block relevant inventory from being returned (see (Diab et al., 2010) on mwe in NLP applications).

The phrases which pass candidate selection are then incorporated into the existing query expansions (i.e. token-to-token mappings, category mappings, attribute mappings). The phrases are a new type of token-to-token mapping which require the query tokens to appear in order and adjacent, i.e. as a mwe phrase, or to be dropped.

2.2 Phrase Candidate Selection

The first stage of the algorithm is candidate selection: from all the possible buyer query n-grams we determine which are potential mwe phrase candidates. We use a straight-forward selection technique in order to gather a large candidate set; at this stage we are concerned with recall, not precision, of the phrases.

First consider required phrases. For a given site (US and Germany here), we consider all the bi- and tri-grams seen in buyer queries. Since e-commerce queries are relatively short, even shorter than web queries, we do not consider longer n-grams. The most frequent of these are then considered candidates. Manual inspection of the candidate set shows a variety of mwe semantic types. As expected in the e-commerce domain, these contain primarily nominal mwe: brand names, product types, and measure phrases (see (Ó Séaghdha and Copestake, 2007) on identifying nominal mwe). Multiword verbs are non-existent in buyer queries and relatively few adjectives are candidates (e.g. *navy blue, brand new*).

Next consider dropped phrases. These are stop words specialized to the e-commerce domain. They are mined from behavioral logs by looking at query-to-query transitions. We consider query transitions where buyers drop a word or phrase in the transition and show increased engagement after the transition. For example, buyers issue the query *used cars for sale* followed by the query *used cars* and subsequently engage with the search results (e.g. view or purchase items). The most frequent n-grams identified by this approach are candidates for dropped phrases and are contextually dropped, i.e. they are dropped when they are parts of specific larger phrases. Query context is important because *for sale* should not be dropped when part of the larger phrase *plastic for sale signs*.

2.3 Phrase Selection: Sorry Mickey

Once we have candidate phrases, we use buyer behavioral data (Carterette et al., 2012) to determine which phrases to require in buyer queries. For each query which contains a given phrase (e.g. for the candidate phrase *apple tv* consider queries such as *apple tv*, *new apple tv*, *apple tv remote*) we see which items were purchased. Item titles from purchased items which contain the phrase are referred to as "phrase bought" while item titles shown in searches are "phrase impressed". We are interested only in high confidence phrases and so focus on purchase behavior: this signal is relatively sparse but is the strongest indicator of buyer interest. To determine the candidates, we want to compute the conditional probability of an item being bought (B(ought)) given a phrase (Ph(rase)).

$$P(B|Ph) = \frac{P(Ph|B) * P(B)}{P(Ph)} \qquad (1)$$

However, this is computationally intensive in that all items retrieved for a query must be considered. In equation 1, P(Ph|B) is easy to compute since only bought items are considered; P(Ph) can be approximated by the ratio of phrases to non-phrases for bought items; P(B) is a constant and hence can be ignored. So, we use the following two metrics based on these probabilities:

- SALE EFFICIENCY: Probability of phrases in bought items, P(Ph|B) > 95%. Ensures quality and acts as an upper bound for the expected loss (equation 2).
- LIFT: Ensures phrasing has a positive revenue impact and handles presentation bias (equation 3).

First consider sale efficiency:

$$P(Ph|B) = \frac{P(Ph \bigcap B)}{P(B)} = \frac{n(ph\_bought)}{n(bought)} \quad (2)$$

One drawback of sale efficiency P(Ph|B) is data sparsity. There is a high false positive rate in identifying phrases when the frequency of bought items is low since it is hard to distinguish signal from noise with a strict threshold. We used Beta-Binomial smoothing to avoid this (Schuckers, 2003; Agarwal et al., 2009). Conceptually, by incorporating Beta-Binomial smoothing, we model the number of phrases bought as a binomial process and use the Beta distribution, which is its conjugate prior, for smoothing the sale efficiency.

However the sale efficiency as captured by the conditional probability of being bought as a phrase (equation 2) does not take into account the distribution of the phrases in the retrieved set. For example for the phrase *apple tv*, 80% of the impressed items contained the phrase while 99% of the bought items contained the phrase, which makes it an excellent phrase. However, for *mount rushmore* 99% of the impressed items contained the phrase while only 97% of the bought items contained the phrase. This implies that the probability of being bought as a phrase for *mount rushmore* is high because of presentation bias (i.e. the vast majority of token matches contain phrases) and not because the phrase itself is an indicator of relevance. To address the issue of presentation bias in P(Ph|B), we use the following lift metric:

$$\frac{P(Ph|B) - P(Ph)}{P(Ph)} > 0 \quad (3)$$

Lift (equation 3) measures the buyers' tendency to purchase phrase items. For a good phrase this value should be high. For example, for *apple tv* this value is +23.13% while for *mount rushmore* it is −1.8%. We only consider phrases that have a positive lift.

Examples of English phrases for buyer queries include *apple tv, bubble wrap, playstation 3, 4 x 4, tank top, nexus 4, rose gold, 1 gb, hot pack, 20 v, kindle fire, hard rock* and *new balance* and German phrases include *geflochtene schnur* (braided line) and *energiespar regler* (energy-saving controller). These form a disparate semantic set including brand names (*new balance*), product types (*bubble wrap*), and units of measure (*1 gb*).

Consider the phrases which were not selected because a significant percentage of the buyer demand was for items where the tokens appeared either in a different order or not adjacent. These include *golf balls, hard drive* and *mickey mouse*. You might ask, what could possibly be a stronger phrase in American English than *mickey mouse*? Closer examination of the buyer behavioral data shows that many buyers are using queries with the tokens *mickey mouse* to find and purchase *mickey and minnie mouse* items. The introduction of *and minnie* in the item titles breaks the query phrase.

3 Experiment Results

We selected phrase candidates for two sites: The US and Germany. These sites were selected because there was significant query and purchasing data which alleviates data sparsity issues and because the language differences allowed us to test the general applicability of the approach.[2]

We created query assets which contained the existing production assets and modified them to include the required phrases and the dropped phrases. The relative query frequency of required phrases (blue) vs. dropped phrases (red) in each experiment is shown in Figure 2.

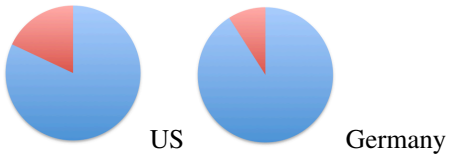

Figure 2: Impacted Query Frequency: red=dropped; blue=required

For US and Germany, 10% of users were ex-

[2]English and German are closely related languages. We plan to apply mwe phrases to Russian and French.

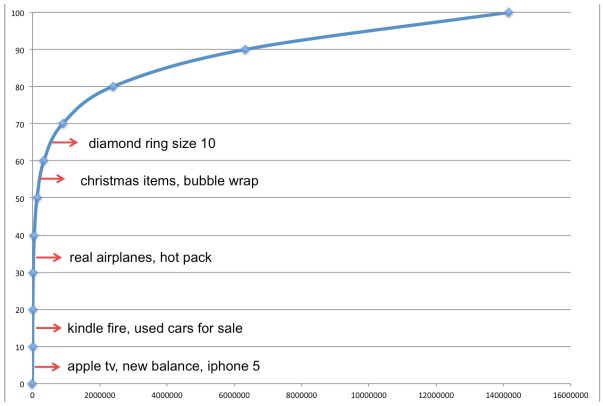

Figure 1: US Phrase Query Impressions: Head-vs.-tail queries

posed to the new phrase assets, while a 10% control[3] were exposed to the existing production assets. The test was run for two weeks. We measured the number of items bought in test vs. control, the revenue, and the behavior of new users. Bought items and revenue are both measured to determine whether changes in purchases are coming from better deals (e.g. bought items might increase while revenue is constant) or improved discovery (e.g. more items are bought at the same price). New user success is measured because new users are generally sensitive to irrelevant items being returned for their queries; the required phrase mwe in this experiment target this use case.

As a result of the phrase experiment, in the US, revenue, bought items, and new user engagement increased statistically significantly ($p<0.1$). The German test showed directionally similar results but was only statistically significant for new buyers. We cannot show proprietary business results, but both experiences are now in production in place of the previous query processing. The graph in Figure 1 shows the distribution of head-vs.-tail queries for the US with some sample affected head queries.

4 Discussion and Conclusion

We described a relatively straight-forward method for detecting phrases in buyer queries. The key insight is that previous buyer purchasing behavior as well as the distribution of phrases in item titles must be used to select which candidate phrases to keep in the final analysis. Many mwe phrases which might be useful in other situations (e.g. our friend *mickey mouse* (§2.3)) are not suitable for buyer queries because many relevant items, as measured by purchases, do not contain these tokens phrases (e.g. *mickey and minnie mouse*).

Among the rejected candidate phrases, the higher confidence ones are likely to be suitable for ranking of the results even though they could not be used to filter out results. This is an area of active research: what mwe phrases can improve the ranking of e-commerce results, especially given the presence of the phrase in the buyer query? Another method to increase phrase coverage is to consider contextualized phrases, whereby token sequences may be a phrase in one query but not in another.

The experiments here were conducted on two of our largest sites, thereby avoiding data sparsity issues. We have used the same algorithm on smaller sites such as Australia: the resulting required phrases and dropped phrases look reasonable but have not been tested experimentally. An interesting question is whether phrases from same-language sites (e.g. UK, Australia, Canada, US) can be combined or whether a site with more behavioral data can be used to learn phrases for smaller sites. The later has been done for Canada using US data.

In sum, mwe phrases improved eBay e-commerce, but it was important to use domain-specific data in choosing the relevant phrases. This suggests that the utility of universal vs. domain specific mwe is an area requiring investigation.

[3]Technically there were two 5% controls which were compared to determine variability within the control group.

References

Deepak Agarwal, Bee-Chung Chen, and Pradheep Elango. 2009. Spatio-temporal models for estimating click-through rate. In *Proceedings of the 18th International Conference on World Wide Web*. ACM.

Ben Carterette, Evangelos Kanoulas, Paul Clough, and Mark Sanderson, editors. 2012. *Information Retrieval Over Query Sessions*. Springer Lecture Notes in Computer Science.

Mona Diab, Valia Kordoni, and Hans Uszkoreit. 2010. Multiword expressions: From theory to applications. Panel at MWE2010.

Ravi Chandra Jammalamadaka and Vamsi Salaka. 2012. Synonym mining and usage in e-commerce. Presented at ECIR.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Diarmuid Ó Séaghdha and Ann Copestake. 2007. Co-occurrence contexts for noun compound interpretation. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 57–64. Association for Computational Linguistics.

Carlos Ramisch, Paulo Schreiner, Marco Idiart, and Aline Villavicencio. 2008. An evaluation of methods for the extraction of multiword expressions. In *Towards a Shared Task for Multiword Expressions*, pages 50–53.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, pages 1–15. Springer-Verlag.

Michael E. Schuckers. 2003. Using the beta-binomial distribution to assess performance of a biometric identification device. *International Journal of Image and Graphics*, pages 523–529.