Can Click Patterns across User's Query Logs Predict Answers to Definition Questions?

Alejandro Figueroa
Yahoo! Research Latin America
Blanco Encalada 2120, Santiago, Chile
afiguero@yahoo-inc.com

Abstract

In this paper, we examined click patterns produced by users of Yahoo! search engine when prompting definition questions. Regularities across these click patterns are then utilized for constructing a large and heterogeneous training corpus for answer ranking. In a nutshell, answers are extracted from clicked web-snippets originating from any class of web-site, including Knowledge Bases (KBs). On the other hand, non-answers are acquired from redundant pieces of text across web-snippets.

The effectiveness of this corpus was assessed via training two state-of-the-art models, wherewith answers to unseen queries were distinguished. These testing queries were also submitted by search engine users, and their answer candidates were taken from their respective returned web-snippets. This corpus helped both techniques to finish with an accuracy higher than 70%, and to predict over 85% of the answers clicked by users. In particular, our results underline the importance of non-KB training data.

1 Introduction

It is a well-known fact that definition queries are very popular across users of commercial search engines (Rose and Levinson, 2004). The essential characteristic of definition questions is their aim for discovering as much as possible descriptive information about the concept being defined (a.k.a. definiendum, pl. definienda). Some examples of this kind of query include "*Who is Benjamin Millepied?*" and "*Tell me about Bank of America*".

It is a standard practice of definition question answering (QA) systems to mine KBs (e.g., online encyclopedias and dictionaries) for reliable descriptive information on the definiendum (Sacaleanu et al., 2008). Normally, these pieces of information (i.e., nuggets) explain different facets of the definiendum (e.g., "*ballet choreographer*" and "*born in Bordeaux*"), and the main idea consists in projecting the acquired nuggets into the set of answer candidates afterwards. However, the performance of this category of method falls into sharp decline whenever few or no coverage is found across KBs (Zhang et al., 2005; Han et al., 2006). Put differently, this technique usually succeeds in discovering the most relevant facts about the most promiment sense of the definiendum. But it often misses many pertinent nuggets, especially those that can be paraphrased in several ways; and/or those regarding ancillary senses of the definiendum, which are hardly found in KBs.

As a means of dealing with this, current strategies try to construct general definition models inferred from a collection of definitions coming from the Internet or KBs (Androutsopoulos and Galanis, 2005; Xu et al., 2005; Han et al., 2006). To a great extent, models exploiting non-KB sources demand considerable annotation efforts, or when the data is obtained automatically, they benefit from empirical thresholds that ensure a certain degree of similarity to an array of KB articles. These thesholds attempt to trade-off the cleanness of the training material against its coverage. Moreover, gathering negative samples is also hard as it is not easy to find wide-coverage authoritative sources of non-descriptive information about a particular definiendum.

Our approach has different innovative aspects

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 99–108,
Avignon, France, April 23 - 27 2012.

compared to other research in the area of definition extraction. It is at the crossroads of query log analysis and QA systems. We study the click behavior of search engines' users with regard to definition questions. Based on this study, we propose a novel way of acquiring large-scale and heterogeneous training material for this task, which consists of:

- automatically obtaining positive samples in accordance with click patterns of search engine users. This aids in harvesting a host of descriptions from non-KB sources in conjunction with descriptive information from KBs.
- automatically acquiring negative data in consonance with redundancy patterns across snippets displayed within search engine results when processing definition queries.

In brief, our experiments reveal that these patterns can be effectively exploited for devising efficient models.

Given the huge amount of amassed data, we additionally contrast the performance of systems built on top of samples originated solely from KB, non-KB, and both combined. Our comparison corroborates that KBs yield massive trustworthy descriptive knowledge, but they do not bear enough diversity to discriminate all answering nuggets within any kind of text. Essentially, our experiments unveil that non-KB data is richer and therefore it is useful for discovering more descriptive nuggets than KB material. But its usage relies on its cleanness and on a negative set. Many people had these intuitions before, but to the best of our knowledge, we provide the first empirical confirmation and quantification.

The road-map of this paper is as follows: section 2 touches on related works; section 3 digs deeper into click patterns for definition questions, subsequently section 4 explains our corpus construction strategy; section 5 describes our experiments, and section 6 draws final conclusions.

2 Related Work

In recent years, definition QA systems have shown a trend towards the utilization of several discriminant and statistical learning techniques (Androutsopoulos and Galanis, 2005; Chen et al., 2006; Han et al., 2006; Fahmi and Bouma, 2006; Katz et al., 2007; Westerhout, 2009; Navigli and Velardi, 2010). Due to training, there is a pressing necessity for large-scale authoritative sources of descriptive and non-descriptive nuggets. In the same manner, there is a growing importance of strategies capable of extracting trustworthy and negative/positive samples from any type of text. Conventionally, these methods interpret descriptions as positive examples, whereas contexts providing non-descriptive information as negative elements. Four representative techniques are:

- centroid vector (Xu et al., 2003; Cui et al., 2004) collects an array of articles about the definiendum from a battery of predetermined KBs. These articles are then used to learn a vector of word frequencies, wherewith answer candidates are rated afterwards. Sometimes web-snippets together with a query reformulation method are exploited instead of pre-defined KBs (Chen et al., 2006).
- (Androutsopoulos and Galanis, 2005) gathered articles from KBs to score 250-characters windows carrying the definiendum. These windows were taken from the Internet, and accordingly, highly similar windows were interpreted as positive examples, while highly dissimilar as negative samples. For this purpose, two thresholds are used, which ensure the trustworthiness of both sets. However, they also cause the sets to be less diverse as not all definienda are widely covered across KBs. Indeed, many facets outlined within the 250-characters windows will not be detected.
- (Xu et al., 2005) manually labeled samples taken from an Intranet. Manual annotations are constrained to a small amount of examples, because it requires substantial human efforts to tag a large corpus, and disagreements between annotators are not uncommon.
- (Figueroa and Atkinson, 2009) capitalized on abstracts supplied by Wikipedia for building language models (LMs), thus there was no need for a negative set.

Our contribution is a novel technique for obtaining heterogeneous training material for defi-

nitional QA, that is to say, massive examples harvested from KBs and non-KBs. Fundamentally, positive examples are extracted from web snippets grounded on click patterns of users of a search engine, whereas the negative collection is acquired via redundancy patterns across web-snippets displayed to the user by the search engine. This data is capitalized by two state-of-the-art definition extractors, which are different in nature. In addition, our paper discusses the effect on the performance of different sorts (KBs and non-KBs) and amount of training data.

As for user clicks, they provide valuable relevance feedback for a variety of tasks, cf. (Radlinski et al., 2010). For instance, (Ji et al., 2009) extracted relevance information from clicked and non-clicked documents within aggregated search sessions. They modelled sequences of clicks as a means of learning to globally rank the relative relevance of all documents with respect to a given query. (Xu et al., 2010) improved the quality of training material for learning to rank approaches via predicting labels using clickthrough data. In our work, we combine click patterns across Yahoo! search query logs with QA techniques to build one-sided and two-sided classifiers for recognizing answers to definition questions.

3 User Click Analysis for Definition QA

In this section, we examine a collection of queries submitted to Yahoo! search engine during the period from December 2010 to March 2011. More specifically, for this analysis, we considered a log encompassing a random sample of 69,845,262 (23,360,089 distinct) queries. Basically, this log comprises the query sent by the user in conjunction with the displayed URLs and the information about the sequence of their clicks.

In the first place, we associate each query with a category in the taxonomy proposed by (Rose and Levinson, 2004), and in this way definition queries are selected. Secondly, we investigate user click patterns observed across these filtered definition questions.

3.1 Finding Definition Queries

According to (Broder, 2002; Lee et al., 2005; Dupret and Piwowarski, 2008), the intention of the user falls into at least two categories: navigational (e.g., "*google*") and informational (e.g., "*maximum entropy models*"). The former entails the desire of going to a specific site that the user has in mind, and the latter regards the goal of learning something by reading or viewing some content (Rose and Levinson, 2004). Navigational queries are hence of less relevance to definition questions, and for this reason, these were removed in congruence with the next three criteria:

- (Lee et al., 2005) pointed out that users will only visit the web site they bear in mind, when prompting navigational queries. Thus, these queries are characterized by clicking the same URL almost all the time (Lee et al., 2005). More precisely, we discarded queries that: a) appear more than four times in the query log; and which at the same time b) its most clicked URL represents more than 98% of all its clicks. Following the same idea, we additionally eliminated prompted URLs and queries where the clicked URL is of the form "www.search-query-without-spaces."

- By the same token, queries containing keywords such as "*homepage*", "*on-line*", and "*sign in*" were also removed.

- After the previous steps, many navigational queries (e.g., "*facebook*") still remained in the query log. We noticed that a substantial portion was signaled by several frequently and indistinctly clicked URLs. Take for instance "*facebook*": "www.facebook.com" and "www.facebook.com/login.php".

  With this in mind, we discarded entries embodied in a manually compiled black list. This list contains the 600 highest frequent cases.

A third category in (Rose and Levinson, 2004) regards resource queries, which we distinguished via keywords like "*image*", "*lyrics*" and "*maps*". Altogether, an amount of (35.67%) 24,916,610 (3,576,817 distinct) queries were seen as navigational and resource. Note that in (Rose and Levinson, 2004) both classes encompassed between 37%-38% of their query set.

Subsequently, we profited from the remaining 44,928,652 (informational) entries for detecting queries where the intention of the user is finding descriptive information about a topic (i.e., definiendum). In the taxonomy delineated by

(Rose and Levinson, 2004), informational queries are sub-categorized into five groups including list, locate, and definitional (directed and undirected). In practice, we filtered definition questions as follows:

1. We exploited an array of expressions that are commonly utilized in query analysis for classifying definition questions (Figueroa, 2010). E.g., "*Who is/was...*", "*What is/was a/an...*", "*define...*", and "*describe...*". Overall, these rules assisted in selecting 332,227 entries.

2. As stated in (Dupret and Piwowarski, 2008), informational queries are typified by the user clicking several documents. In light of that, we say that some definitional queries are characterized by multiple clicks, where at least one belongs to a KB. This aids in capturing the intention of the user when looking for descriptive knowledge and only entering noun phrases like "*thoracic outlet syndrome*":

www.medicinenet.com **en.wikipedia.org** health.yahoo.net www.livestrong.com
health.yahoo.net **en.wikipedia.org**
www.medicinenet.com www.mayoclinic.com **en.wikipedia.org**
www.nismat.org **en.wikipedia.org**

Table 1: Four distinct sequences of hosts clicked by users given the search query: "*thoracic outlet syndrome*".

   In so doing, we manually compiled a list of 36 frequently clicked KB hosts (e.g., Wikipedia and Britannica encyclopedia). This filter produced 567,986 queries.

Unfortunately, since query logs stored by search engines are not publicly available due to privacy and legal concerns, there is no accessible training material to build models on top of annotated data. Thus, we exploited the aforementioned hand-crafted rules to connect queries to their respective category in this taxonomy.

3.2 User Click Patterns

In substance, the first filter recognizes the intention of the user by means of the formulation given by the user (e.g., "*What is a/the/an...*"). With regard to this filter, some interesting observations are as follows:

- In 40.27% of the entries, users did not visit any of the displayed web-sites. Consequently, we concluded that the information conveyed within the multiple snippets was often enough to answer the respective definition question. In other words, a significant fraction of the users were satisfied with a small set of brief, but quickly generated descriptions.

- In 2.18% of these cases, the search engine returned no results, and a few times users tried another paraphrase or query, due to useless results or misspellings.

- We also noticed that definition questions matched by these expressions are seldom related to more than one click, although informational queries produce several clicks, in general. In 46.44% of the cases, the user clicked a sole document, and more surprisingly, we observed that users are likely to click sources different from KBs, in contrast to the widespread belief in definition QA research. Users pick hits originating from small but domain-specific web-sites as a result of at least two effects: a) they are looking for minor or ancillary senses of the definiendum (e.g., "*ETA*" in "www.travel-industry-dictionary.com"); and more pertinent b) the user does not trust the information yielded by KBs and chooses more authoritative resources, for instance, when looking for reliable medical information (e.g., "*What is hypothyroidism?*", and "*What is mrsa infection?*").

While the first filter infers the intention of the user from the query itself, the second deduces it from the origin of the clicked documents. With regard to this second filter, clicking patterns are more disperse. Here, the first two clicks normally correspond to the top two/three ranked hits returned by the search engine, see also (Ji et al., 2009). Also, sequences of clicks signal that users

normally visit only one site belonging to a KB, and at least one coming from a non-KB (see Table 1).

All in all, the insight gained in this analysis allows the construction of an heterogeneous corpus for definition question answering. Put differently, these user click patterns offer a way to obtain huge amounts of heterogeneous training material. In this way the heavy dependence of open-domain description identifiers on KB data can be alleviated.

4 Click-Based Corpus Acquisition

Since queries obtained by the previous two filters are not associated with the actual snippets seen by the users (due to storage limitations), snippets were recovered by means of submitting the queries to Yahoo! search engine.

After retrieval, we benefited from OpenNLP[1] for detecting sentence boundaries, tokenization and part-of-speech (POS) information. Here, we additionally interpreted truncations ("...") as sentence delimiters. POS tags were used to recognize and replace numbers with a placeholder (#CD#) as a means of creating sentence templates. We modified numbers as their value is just as often confusing as useful (Baeza-Yates and Ribeiro-Neto, 1999).

Along with numbers, sequences of full and partial matches of the definiendum were also substituted with placeholders, "#Q#" and "#QT#", respectively. To exemplify, consider this pre-processed snippet regarding "*Benjamin Millepied*" from "www.mashceleb.com":

```
#Q# / News &amp;amp; Biography - MashCeleb
Latest news coverage of #Q#
#Q# ( born #CD# ) is a principal dancer
at New York City Ballet and a ballet
choreographer...
```

We benefit from these templates for building both a positive and a negative training set.

4.1 Negative Set

The negative set comprised templates appearing across all (clicked and unclicked) web-snippets, which at the same time, are related to more than five distinct queries. We hypothesize that these prominent elements correspond to non-informative, and thus non-descriptive, content as they appear within snippets across several questions. In other words: "*If it seems to answer every question, it will probably answer no question*". Take for instance:

```
Information about #Q# in the Columbia
Encyclopedia , Computer Desktop
Encyclopedia , computing dictionary
```

Conversely, templates that are more plausible to be answers are strongly related to their specific definition questions, and consequently, they are low in frequency and unlikely to be in the result set of a large number of queries. This negative set was expanded with templates coming from titles of snippets, which at the same time, have a frequency higher than four across all snippets (independent on which queries they appear). This process cooperated on gathering 1,021,571 different negative examples. In order to measure the precision of this process, we randomly selected and checked 1,000 elements, and we found an error of 1.3%.

4.2 Positive Set

As for the positive set, this was constructed only from the summary section of web-snippets clicked by the users. We constrained these snippets to bear a title template associated with at least two web-snippets clicked for two distinct queries. Some good examples are:

```
What is #Q# ?  Choices and Consequences.
Biology question :  What is an #Q# ?
```

Since clicks are linked with entire snippets, it is uncertain which sentences are genuine descriptions (see the previous example). Therefore, we removed those templates already contained in the negative set, along with those samples that matched an array of well-known handcrafted rules. This set included:

a. sentences containing words such as "*ask*", "*report*", "*say*", and "*unless*" (Kil et al., 2005; Schlaefer et al., 2007);

b. sentences bearing several named entities (Schlaefer et al., 2006; Schlaefer et al., 2007), which were recognized by the number of tokens starting with a capital letter versus those starting with a lowercase letter;

c. statements of persons (Schlaefer et al., 2007); and

[1] http://opennlp.sourceforge.net

d. we also profited from about five hundred common expressions across web snippets including "*Picture of*", and "*Jump to : navigation , search*", as well as "*Recent posts*".

This process assisted in acquiring 881,726 different examples, where 673,548 came from KBs. Here, we also randomly selected 1,000 instances and manually checked if they were actual descriptions. The error of this set was 12.2%.

To put things into perspective, in contrast to other corpus acquisition approaches, the present method generated more than 1,800,000 positive and negative training samples combined, while the open-domain strategy of (Miliaraki and Androutsopoulos, 2004; Androutsopoulos and Galanis, 2005) ca. 20,000 examples, the close-domain technique of (Xu et al., 2005) about 3,000 and (Fahmi and Bouma, 2006) ca. 2,000.

5 Answering New Definition Queries

In our experiments, we checked the effectiveness of our user click-based corpus acquisition technique by studying its impact on two state-of-the-art systems. The first one is based on the **bi-term LMs** proposed by (Chen et al., 2006). This system requires only positive samples as training material. Conversely, our second system capitalizes on both positive and negative examples, and it is based on the **Maximum Entropy** (ME) models presented by (Fahmi and Bouma, 2006). These ME[2] models amalgamated bigrams and unigrams as well as two additional syntactic features, which were not applicable to our task (i.e, sentence position). We added to this model the sentence length as a feature in order to homologate the attributes used by both systems, therefore offering a good framework to assess the impact of our negative set. Note that (Fahmi and Bouma, 2006), unlike us, applied their models only to sentences observing some specific syntactic patterns.

With regard to the **test set**, this was constructed by manually annotating 113,184 sentence templates corresponding to 3,162 unseen definienda. In total, this array of unseen testing instances encompassed 11,566 different positive samples. In order to build a balanced testing collection, the same number of negative examples were randomly selected. Overall, our testing set contains 23,132 elements, and some illustrative annotations are shown in Table 2. It is worth highlighting that these examples signal that our models are considering pattern-free descriptions, that is to say, unlike other systems (Xu et al., 2003; Katz et al., 2004; Fernandes, 2004; Feng et al., 2006; Figueroa and Atkinson, 2009; Westerhout, 2009) which consider definitions aligning an array of well-known patterns (e.g., "*is a*" and "*also known as*"), our models disregard any class of syntactic constraint.

As to a **baseline** system, we accounted for the **centroid vector** (Xu et al., 2003; Cui et al., 2004). When implementing, we followed the blueprint in (Chen et al., 2006), and it was built for each *definiendum* from a maximum of 330 web snippets fetched by means of Bing Search. This baseline achieved a modest performance as it correctly classified 43.75% of the testing examples. In detail, 47.75% out of the 56.25% of the misclassified elements were a result of data-sparseness. This baseline has been widely used as a starting point for comparison purposes, however it is hard for this technique to discover diverse descriptive nuggets. This problem stems from the narrow-coverage of the centroid vector learned for the respective definienda (Zhang et al., 2005). In short, these figures support the necessity for more robust methods based on massive training material.

Experiments. We trained both models by systematically increasing the size of the training material by 1%. For this, we randomly split the training data into 100 equally sized packs, and systematically added one to the previously selected sets (i. e., 1%, 2%, 3%, . . ., 99%, 100%). We also experimented with: 1) positive examples originated solely from KBs; 2) positive samples harvested only from non-KBs; and eventually 3) all positive examples combined.

Figure 1 juxtaposes the outcomes accomplished by both techniques under the different configurations. These figures, compared with results obtained by the baseline, indicate the important contribution of our corpus to tackle data-sparseness. This contrast substantiates our claim that click patterns can be utilized as indicators of answers to definition questions. Since our models ignore definition patterns, they have the potential of detecting a wide diversity of descriptive information.

Further, the improvement of about 9%-10% by

[2]http://maxent.sourceforge.net/about.html

Label	Example/Template
+	Propylene #Q# is a type of alcohol made from fermented yeast and carbohydrates and is commonly used in a wide variety of products .
+	#Q# is aggressive behavior intended to achieve a goal .
+	In Hispanic culture , when a girl turns #CD# , a celebration is held called the #Q#, symbolizing the girl 's passage to womanhood .
+	Kirschwasser , German for " cherry water " and often shortened to #Q# in English-speaking countries , is a colorless brandy made from black ...
+	From the Gaelic 'dubhglas ' meaning #Q#, #QT# stream , or from the #QT# river .
+	Council Bluffs Orthopedic Surgeon Doctors physician directory - Read about #Q#, damage to any of the #CD# tendons that stabilize the shoulder joint .
+	It also occurs naturally in our bodies in fact , an average size adult manufactures up to #CD# grams of #Q# daily during normal metabolism .
-	Sterling Silver #Q# Hoop Earrings Overstockjeweler.com
-	I know V is the rate of reaction and the #Q# is hal ...
-	As sad and mean as that sounds , there is some truth to it , as #QT# as age their bodies do not function as well as they used to (in all respects) so there is a ...
-	If you 're new to the idea of Christian #Q#, what I call " the wild things of God ,
-	A look at the Biblical doctrine of the #QT# , showing the biblical basis for the teaching and including a discussion of some of the common objections .
-	#QT# is Users Choice (application need to be run at #QT# , but is not system critical) , this page shows you how it affects your Windows operating system .
-	Your doctor may recommend that you use certain drugs to help you control your #Q# .
-	Find out what is the full meaning of #Q# on Abbreviations.com !

Table 2: Samples of manual annotations (testing set).

means of exploiting our negative set makes its positive contribution clear. In particular, this supports our hypothesis that redundancy across web-snippets pertaining to several definition questions can be exploited as negative evidence. On the whole, this enhancement also suggests that ME models are a better option than LMs.

Furthermore, in the case of ME models, putting together evidence from KB and non-KBs betters the performance. Conversely, in the case of LMs, we do not observe a noticeable improvement when unifying both sources. We attribute this difference to the fact that non-KB data is noisier, and thus negative examples are necessary to cushion this noise. By and large, the outcomes show that the usage of descriptive information derived exclusively from KBs is not the best, but a cost-efficient solution.

Incidentally, Figure 1 reveals that more training data does not always imply better results. Overall, the best performance (ME-combined → 80.72%) was reaped when considering solely 32% of the training material. Hence, ME-KB finished with the best performance when accounting for about 215,500 positive examples (see Table 3). Adding more examples brought about a decline in accuracy. Nevertheless, this fraction (32%) is still larger than the data-sets considered by other open-domain Machine Learning approaches (Miliaraki and Androutsopoulos, 2004; Androutsopoulos and Galanis, 2005).

Best Conf. of	Accuracy	True positives	Positive examples
ME-combined	80.72%	88%	881,726
ME-KB	80.33%	89.37%	673,548
ME-N-KB	78.99%	93.38%	208,178

Table 3: Comparison of performance, the total amount and origin of training data, and the number of recognized descriptions.

In detail, when contrasting the confusion matrices of the best configurations accomplished by ME-combined (80.72%), ME-KB (80.33%) and ME-N-KB (78.99%), one can find that ME-combined correctly identified 88% of the answers (true positives), while ME-KB 89.37% and ME-N-KB 93.38% (see Table 3).

Interestingly enough, non-KB data only embodies 23.61% of all positive training material, but it still has the ability to recognize more answers. Despite of that, the other two strategies outperform ME-N-KB, because they are able

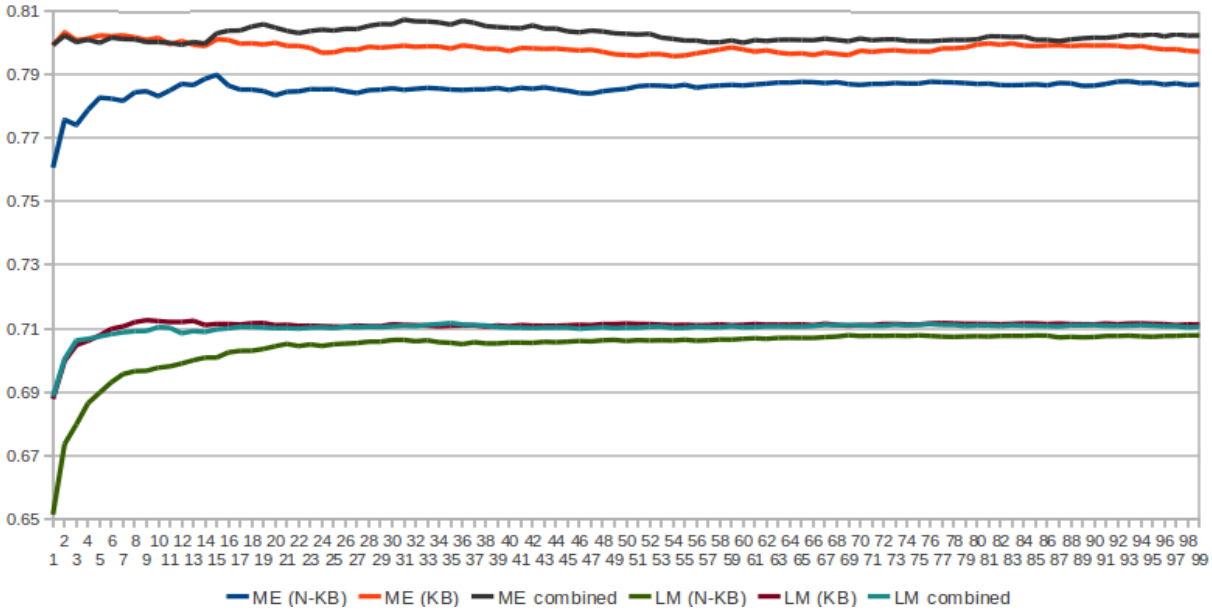

Figure 1: Results for each configuration (accuracy).

to correctly label more negative test examples. Given these figures, we can conclude that this is achieved by mitigating the impact of the noise in the training corpus by means of cleaner (KB) data.

We verified this synergy by inspecting the number of answers from non-KBs detected by the three top configurations in Table 3: ME-combined (9,086), ME-KB (9,230) and ME-N-KB (9,677). In like manner, we examined the confusion matrix for the best configuration (ME-combined → 80.72%): 1,388 (6%) positive examples were mislabeled as negative, while 3,071 (13.28%) negative samples were mistagged as positive.

In addition, we performed significance tests utilizing two-tailed paired t-test at 95% confidence interval on twenty samples. For this, we used only the top three configurations in Table 3 and each sample was determined by using boostrapping resampling. Each sample has the same size of the original test corpus. Overall, the tests implied that all pairs were statistically different from each other.

In summary, the results show that both negative examples and combining positive examples from heterogeneous sources are indispensable to tackle any class of text. However, it is vital to lessen the noise in non-KB data, since this causes a more adverse effect on the performance. Given the upperbound in accuracy, our outcomes indicate that cleanness and quality are more important than the size of the corpus. Our figures additionally suggest that more effort should go into increasing diversity than the number of training instances. In light of these observations, we also conjecture that a more reduced, but diverse and manually annotated, corpus might be more effective. In particular, a manually checked corpus distilled by inspecting click patterns across query logs of search engines.

Lastly, in order to evaluate how good a click predictor the three top ME-configurations are, we focused our attention only on the manually labeled positive samples (answers) that were clicked by the users. Overall, 86.33% (ME-combined), 88.85% (ME-KB) and 92.45% (ME-N-KB) of these responses were correctly predicted. In light of that, one can conclude that (clicked and non-clicked) answers to definition questions can be identified/predicted on the basis of user's click patterns across query logs.

From the viewpoint of search engines, web snippets are computed off-line, in general. In so doing, some methods select the spans of text bearing query terms with the potential of putting the document on top of the rank (Turpin et al., 2007; Tsegay et al., 2009). This helps to create an abridged version of the document that can quickly produce the snippet. This has to do with the trade-off between storage capacity, indexing, and retrieval speed. Ergo, our technique can help to de-

termine whether or not a span of text is worth expanding, or in some cases whether or not it should be included in the snippet view of the document. In our instructive snippet, we now might have:

```
Benjamin Millepied / News &amp;
Biography - MashCeleb
Benjamin Millepied (born 1977) is a
principal dancer at New York City Ballet
and a ballet choreographer of
international reputation.  Millepied was
born in Bordeaux, France.  His...
```

Improving the results of informational (e.g., definition) queries, especially of less frequent ones, is key for competing commercial search engines as they are embodied in the non-navigational tail where these engines differ the most (Zaragoza et al., 2010).

6 Conclusions

This work investigates into the click behavior of commercial search engine users regarding definition questions. These behaviour patterns are then exploited as a corpus acquisition technique for definition QA, which offers the advantage of encompassing positive samples from heterogoneous sources. In contrast, negative examples are obtained in conformity to redundancy patterns across snippets, which are returned by the search engine when processing several definition queries. The effectiveness of these patterns, and hence of the obtained corpus, was tested by means of two models different in nature, where both were capable of achieving an accuracy higher than 70%.

As a future work, we envision that answers detected by our strategy can aid in determining some query expansion terms, and thus to devise some relevance feedback methods that can bring about an improvement in terms of the recall of answers. Along the same lines, it can cooperate on the visualization of the results by highlighting and/or extending truncated answers, that is more informative snippets, which is one of the holy grail of search operators, especially when processing informational queries.

NLP tools (e.g., parsers and name entity recognizers) can also be exploited for designing better training data filters and more discriminative features for our models that can assist in enhancing the performance, cf. (Surdeanu et al., 2008; Figueroa, 2010; Surdeanu et al., 2011). However, this implies that these tools have to be re-trained to cope with web-snippets.

Acknowledgements

This work was partially supported by R&D project FONDEF D09I1185. We also thank our reviewers for their interesting comments, which helped us to make this work better.

References

I. Androutsopoulos and D. Galanis. 2005. A practically Unsupervised Learning Method to Identify Single-Snippet Answers to Definition Questions on the web. In *HLT/EMNLP*, pages 323–330.

R. Baeza-Yates and B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison Wesley.

A. Broder. 2002. A Taxonomy of Web Search. *SIGIR Forum*, 36:3–10, September.

Y. Chen, M. Zhon, and S. Wang. 2006. Reranking Answers for Definitional QA Using Language Modeling. In *Coling/ACL-2006*, pages 1081–1088.

H. Cui, K. Li, R. Sun, T.-S. Chua, and M.-Y. Kan. 2004. National University of Singapore at the TREC 13 Question Answering Main Task. In *Proceedings of TREC 2004*. NIST.

Georges E. Dupret and Benjamin Piwowarski. 2008. A user browsing model to predict search engine click data from past observations. In *SIGIR '08*, pages 331–338.

Ismail Fahmi and Gosse Bouma. 2006. Learning to Identify Definitions using Syntactic Features. In *Proceedings of the Workshop on Learning Structured Information in Natural Language Applications*.

Donghui Feng, Deepak Ravichandran, and Eduard H. Hovy. 2006. Mining and Re-ranking for Answering Biographical Queries on the Web. In *AAAI*.

Aaron Fernandes. 2004. Answering Definitional Questions before they are Asked. Master's thesis, Massachusetts Institute of Technology.

A. Figueroa and J. Atkinson. 2009. Using Dependency Paths For Answering Definition Questions on The Web. In *WEBIST 2009*, pages 643–650.

Alejandro Figueroa. 2010. *Finding Answers to Definition Questions on the Web*. Phd-thesis, Universitaet des Saarlandes, 7.

K. Han, Y. Song, and H. Rim. 2006. Probabilistic Model for Definitional Question Answering. In *Proceedings of SIGIR 2006*, pages 212–219.

Shihao Ji, Ke Zhou, Ciya Liao, Zhaohui Zheng, Gui-Rong Xue, Olivier Chapelle, Gordon Sun, and Hongyuan Zha. 2009. Global ranking by exploiting user clicks. In *Proceedings of the 32nd international ACM SIGIR conference on Research and*

development in information retrieval, SIGIR '09, pages 35–42, New York, NY, USA. ACM.

B. Katz, M. Bilotti, S. Felshin, A. Fernandes, W. Hildebrandt, R. Katzir, J. Lin, D. Loreto, G. Marton, F. Mora, and O. Uzuner. 2004. Answering multiple questions on a topic from heterogeneous resources. In *Proceedings of TREC 2004*. NIST.

B. Katz, S. Felshin, G. Marton, F. Mora, Y. K. Shen, G. Zaccak, A. Ammar, E. Eisner, A. Turgut, and L. Brown Westrick. 2007. CSAIL at TREC 2007 Question Answering. In *Proceedings of TREC 2007*. NIST.

Jae Hong Kil, Levon Lloyd, and Steven Skiena. 2005. Question Answering with Lydia (TREC 2005 QA track). In *Proceedings of TREC 2005*. NIST.

U. Lee, Z. Liu, and J. Cho. 2005. Automatic Identification of User Goals in Web Search. In *Proceedings of the 14th WWW conference*, WWW '05, pages 391–400.

S. Miliaraki and I. Androutsopoulos. 2004. Learning to identify single-snippet answers to definition questions. In *COLING '04*, pages 1360–1366.

Roberto Navigli and Paola Velardi. 2010. LearningWord-Class Lattices for Definition and Hypernym Extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*.

Filip Radlinski, Martin Szummer, and Nick Craswell. 2010. Inferring query intent from reformulations and clicks. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 1171–1172, New York, NY, USA. ACM.

Daniel E. Rose and Danny Levinson. 2004. Understanding User Goals in Web Search. In *WWW*, pages 13–19.

B. Sacaleanu, G. Neumann, and C. Spurk. 2008. DFKI-LT at QA@CLEF 2008. In *In Working Notes for the CLEF 2008 Workshop*.

Nico Schlaefer, P. Gieselmann, and Guido Sautter. 2006. The Ephyra QA System at TREC 2006. In *Proceedings of TREC 2006*. NIST.

Nico Schlaefer, Jeongwoo Ko, Justin Betteridge, Guido Sautter, Manas Pathak, and Eric Nyberg. 2007. Semantic Extensions of the Ephyra QA System for TREC 2007. In *Proceedings of TREC 2007*. NIST.

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2008. Learning to Rank Answers on Large Online QA Collections. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL 2008)*, pages 719–727.

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2011. Learning to rank answers to non-factoid questions from web collections. *Computational Linguistics*, 37:351–383.

Yohannes Tsegay, Simon J. Puglisi, Andrew Turpin, and Justin Zobel. 2009. Document compaction for efficient query biased snippet generation. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pages 509–520, Berlin, Heidelberg. Springer-Verlag.

Andrew Turpin, Yohannes Tsegay, David Hawking, and Hugh E. Williams. 2007. Fast generation of result snippets in web search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 127–134, New York, NY, USA. ACM.

Eline Westerhout. 2009. Extraction of definitions using grammar-enhanced machine learning. In *Proceedings of the EACL 2009 Student Research Workshop*, pages 88–96.

Jinxi Xu, Ana Licuanan, and Ralph Weischedel. 2003. TREC2003 QA at BBN: Answering Definitional Questions. In *Proceedings of TREC 2003*, pages 98–106. NIST.

J. Xu, Y. Cao, H. Li, and M. Zhao. 2005. Ranking Definitions with Supervised Learning Methods. In *WWW2005*, pages 811–819.

Jingfang Xu, Chuanliang Chen, Gu Xu, Hang Li, and Elbio Renato Torres Abib. 2010. Improving quality of training data for learning to rank using click-through data. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 171–180, New York, NY, USA. ACM.

H. Zaragoza, B. Barla Cambazoglu, and R. Baeza-Yates. 2010. We Search Solved? All Result Rankings the Same? In *Proceedings of CKIM'10*, pages 529–538.

Zhushuo Zhang, Yaqian Zhou, Xuanjing Huang, and Lide Wu. 2005. Answering Definition Questions Using Web Knowledge Bases. In *Proceedings of IJCNLP 2005*, pages 498–506.