Context-Aware Semantic Classification of Search Queries for Browsing Community Question-Answering Archives

Alejandro Figueroa[b,a], Günter Neumann[c]

[a] *Yahoo! Research Latin America, Blanco Encalada 2120, Santiago, Chile*
[b] *Facultad de Ingeniería, Universidad Andrés Bello, Santiago, Chile*
[c] *German Research Center for Artificial Intelligence, DFKI GmbH, Saarbrücken, Germany*

Abstract

Community Question-Answering (cQA) platforms have become massive repositories of user-generated content. To a great extent, these archives have proven to be highly re-usable. For instance, web search engines profit from their best answers for enhancing user experience when resolving question-like queries. Hence, considerable research efforts have gone into trying to revitalize and retrieve past answers contained in these archives. However, similarly to traditional web search, there is a linguistic gap between cQA questions and question-like search queries that are utilized for fetching information from these cQA repositories (e.g., "*rib pain after ovulation*" and "*iron oxide household*").

In fact, this gap does not only consider linguistic features, but also structural and social attributes. On the one hand side, cQA questions are long-winded, they can bear a title and a body, and community members are compelled to categorize questions at posting time. On the other hand side, search queries come as an uncategorized short stream of words. Moreover, in juxtaposition to cQA question, users typically submit streaks of semantically related search queries, when attempting to fulfil their information needs.

This work digs deep into effectively exploiting semantic cues, yielded by preceding queries within the same user session, for classifying question-like search queries into twenty-six semantic cQA question categories. In order to find significant discriminative properties, we carried out experiments on a large-scale dataset acquired automatically. Broadly speaking, our results indicate that more effective semantic features can be computed as long as we account for a larger number of previous queries. In particular, facilitating Explicit Semantic Analysis for modelling the query context shows to be extremely helpful for increasing the classification rate.

Keywords: search query understanding; query classification; search session analysis; user experience; information retrieval; web and text mining; feature analysis; natural language processing;

*Corresponding author; phone: +56 (2) 29784632
Email addresses: alejandro.figueroa@unab.cl (Alejandro Figueroa), neumann@dfki.de (Günter Neumann)

1. Introduction

Community Question Answering (cQA) services such as Yahoo!Answers[1], StackExchange[2] and many others[3] have become extremely popular for maintaining and distributing user-generated content in form of textual questions and their respective answers on a very large scale. Web users take advantage of community Question Answering services for getting help from other individuals, who know or can readily produce satisfactory precise answers, or like in many cases, can provide help by conducting opinion polls and surveys. Due to the intrinsic dynamic of these platforms, posted questions can receive several responses from multiple members, which can not only be supplementary or complementary to each other, but they can also reflect different sentiments and aspects.

Nowadays, the largest cQA services maintain over 100 million answered questions, making them a very huge and valuable repository of knowledge for automatic text analytics and knowledge acquisition [1, 2]. At their core, cQA archives (or cQA Knowledge base, cQA-KB for short) keep each user question linked with all the answers given by the community members (if any). In this scheme, its selected best answer is specially marked. This design provides the foundation for additional features that make cQA-KBs more attractive. For instance, they are usually organized in categories, which are chosen by members when submitting new questions. These categories can then be used for locating contents on topics of interest and question goals [3].

In essence, many cQA platforms are perceived as the synergy of a information-seeking and a social network [4], because members can post any kind of question, either simple, complex, detailed, or questions about opinions. When taking part in this network, members additionally provide social capital: rate the answers' quality (via positive/negative votes, thumbs-up/thumbs-down, etc.) and post comments. Through these social interactions, members share their knowledge so as to construct a valuable, rapidly growing massive archive of questions and answers rated by humans.

Another major feature of most cQA services is their search facility that allows their members for browsing their archives. By doing so, they capitalize on traditional information retrieval approaches such that the community members can formulate and send a sequence of an arbitrary amount of question-like queries to a search box until they find an old question (if any) pertaining to their current need. This sort of approach assists cQA platforms in re-using and revitalizing past questions and answers indexed in their archives. In like manner, web search engines can benefit from this facility for enhancing user experience, whenever they detect that question-like search queries are submitted. As a matter of fact, web engines return hits found by browsing these archives at the top positions of their rank, displaying not only links to strongly related questions, but also producing their snippets from the best answers contained therein. All in all, by dispensing this search facility, cQA platforms aim at reducing the inherent delay time that exists between the moment members post new questions and the arrival of good answers.

[1] `https://answers.yahoo.com/`

[2] `http://stackexchange.com/`

[3] `http://en.wikipedia.org/wiki/List_of_question-and-answer_websites`

However, previous works have shown the existence of a linguistic gap between search queries and web documents [5, 6]. As a natural consequence, this incongruence is also observed between cQA pages and question-like queries. To illustrate, cQA question titles frequently consist of multiple-sentences [7, 8]. In the case of cQA material, this gap is not only linguistic, it also entails structural and social attributes. To be more precise, cQA questions comprise a title and a body, ergo they are long-winded with respect to search queries, which normally encompass few words. Furthermore, cQA questions are categorized by the submitter at posting time in consonance with a taxonomy proposed by the cQA service. Conversely, members are not compelled to categorize question-like search queries when using the search box or when they access it via a web search engine. Note also that there are extra features that widen this gap, for instance questions are normally voted by members of the community, search queries are not. A key advantage of search queries to cQA questions regards the fact that users typically submit streaks of semantically related queries, when seeking to fulfil their information need. On the other hand, community members seldom post streaks of semantically related new questions to the platform.

This work aims at narrowing this gap by categorizing question-like queries in congruence with a semantic taxonomy of questions yield by a cQA service, i.e., Yahoo! Answers in our case. The assumption here is that if we are able to correctly induce the semantic class, we might be able to direct the search more effectively towards good answers contained in these archives, for example, by applying category-specific models. We think that this is a reasonable standpoint, since most search engines are seeking to categorize their documents to enhance search experience. More precisely, our paper focuses on inspecting and effectively exploiting semantic cues found within the context of these question-like queries for improving the classification rate. Here, the context is set up or represented by the series of previous search queries entered by the user during the same session.

As mentioned, search queries do not have a semantic class associated, hence we are using a cQA-KB (based on a Yahoo! Answers archive) as a knowledge base that: a) defines the semantic classes; and b) provide an explicit mapping between search queries and cQA titles/pages (questions and their user-assigned classes). In this way, we learn a model for inducing the semantic class of a new search query by "analogy" of how cQA questions are semantically classified and linked to search queries (via the search engine log). Since, we are learning from the cQA knowledge store and its association to the determined search sessions how to semantically classify search queries, we believe that our research results also contribute towards an effective integration of search engines and cQA KBs.

Our method recognizes question-like queries by inspecting their associations with Yahoo! Answers pages via user clicks, providing the additional benefit of linking each query with an entry in the Yahoo! Answers category system. Thus our target semantic labeling set comprises 26 categories including business, environment, health, pets, sports and travel. As a consequence, we are able to completely automatize our approach without the need of manually annotated training material, and to automatically create a huge annotated corpus of semantically labeled question-like search queries. We then consider all search queries of a current session entered before the current labeled one as candidate sources for contextual information. To be more precise, the contribution of this paper extends our earlier work described in [9] as follows:

- We provide additional experimental evidence that corroborate our earlier finding that as long as the context increases, the classification rate is highly likely to improve, especially when this context is used for harvesting a larger amount of effective features. In so doing, we tested two distinct multi-class discriminant functions (i.e, Maximum Entropy Models (MaxEnt) and Winnow2) on a massive automatically acquired dataset.

- We extract features on the grounds of our early finding that hypernymy and meronymy semantic relations, between terms conveyed in the same session, are informative for this classification task. In this study, we carry out experiments on a substantially wider variety of fine-grained linguistically-oriented semantic attributes. In a nutshell, we discover that benefiting from Explicit Semantic Analysis as a means of inferring semantics cues is extremely helpful for reducing, and thus determining, the semantic range of question-like search queries. To be more exact, our experimental results indicates that the more context we exploit, the shorter the semantic vector should be.

- In order to study the impact of each of these semantic features thoroughly, we compute them by considering different levels of context, i.e., by a different number of preceding queries. In this way, we can find the number of queries necessary for the effective computation of each attribute.

Many people had these intuitions before, but to the best of our knowledge, we undertake the first painstaking large-scale research, providing empirical confirmation and quantification. The remainder of the paper is organized as follows. The next section 2 presents a summary of related scientific work. The corpus acquisition process is dissected in section 3. In section 4 we then describe and motivate the different features, we want to derive from the target search query (the element being classified) and from its previous historical elements in the transaction (query session). The experiments are then described in detail in section 5. Finally, in section 6 we summarize our findings and give a brief overview of future work.

2. Related Work

To the best of our knowledge, our work pioneers the idea of profiting from search sessions for semantically categorizing question-like informational search queries. Broadly speaking, our study is related to search query understanding, session analysis, user-click analysis, and semantic categorization in community question answering.

2.1. Search Query Understanding

In a broad sense, [10] proposed a framework for understanding the underlying goals of user searches. They outlined a taxonomy where the first level models three ends: informational (learn something by reading or viewing), navigational (going to a specific web-site) and resource (obtain videos, maps, etc.).

Later, in a more specific manner, the work of [11] seeks to understand search queries bearing a particular type of entity (e.g., musician) by classifying their generic user intents (e.g., songs, tickets, lyrics and mp3). They built a

taxonomy of search intents by exploiting clustering algorithms, capturing words and phrases that frequently co-occur with entities in user queries, and by examining the click relationships between different intent phrases. Posteriorly, [12] extended this work by organizing query terms within named entity queries into topics, helping to better understand major search intents about entities. The study of [13] presented an unsupervised approach to cluster queries with similar intents that are patterns consisting of a sequence of semantic concepts or lexical items. Recently, [14] has shown the effectiveness of several linguistic features on the three-way classification of user intents.

In effect, named entities cooperate on understanding user intents better, however detecting named entities in search queries is a difficult task, because named entities are not in standard form and search queries are typically very short [15, 16] and their capitalization is inconsistent [5]. Thus, [17] exploited query sequences in search sessions for dealing with the lack of context in short queries, when distinguishing named entities on queries.

2.2. Session Analysis

The segmentation of query logs into short topical sessions is a difficult task, which has to take into account temporal, lexical, and topical clues for identifying proper session boundaries [18]. Basically, a search session is defined as a sequence of queries issued by a single user within a specific time limit [19]. In [20], this perspective is extended by modeling and analyzing complex, multi-session information needs, which they call cross-session search tasks. The work of [21] concluded that the users history of input queries and visited pages are features pertinent to the users current search intent and these can be used to better identify the search intent behind the query.

In terms of informative attributes, [22] tested the performance several distance metrics for comparing pairs of queries (e.g., levenshtein and longest common substring) on extracting tasks from sessions. In this spirit, [23] profited from distance metrics for implementing query-relatedness features in context-aware ranking. Along these lines, [24] also capitalized on similar features in their learning-to-rank approach for judging the relevance of documents in web search. The work of [25] used the recurrence of queries within the search session for extracting tasks. In [26], attributes including query terms, explicit (i.e., Google and Yahoo! Web Directories) and implicit feedbacks along with the direct association of adjacent labels were used for classifying queries according to taxonomy composed of seven level-1 types provided by the ACM KDD Cup 2005.

An important aspect, tackled in our paper, is whether and how much contextual information extracted from user-specific search query sessions helps to effectively train and apply a model to predict the semantic category of a question-like informational search query (cf. [10, 27]). We will show that the classification accuracy improved in congruence with the number of previous queries used to model the question context. A main aspect is to consider the context of an (semantically unclassified) query by means of a sequence of previously submitted queries. Thus, the identification of proper search session boundaries is crucial. Consequently, we perform different experiments in order to explore the effect of different contextual window sizes for the prediction of the semantic class of the search query in question.

2.3. User-Click Analysis for cQA

The analysis of user-clicks for improving search in cQA has been widely studied, especially for the automatic identification of question paraphrases [28]. The core idea is to use the user generated questions of a cQA along with search engine query logs to automatically formulate effective questions or paraphrases in order to improve search in cQA. The works of [1, 29] and others have furthered this idea into the direction of generation of new questions from queries and for paraphrase ranking.

The idea behind [30] is reusing resolved questions for estimating the probability of new questions to be answered by past best answers. Their strategy capitalized on Latent Dirichlet Allocation (LDA) for inferring latent topics for each category, and they compared the distribution of topics for the new and previous questions as well as the answers. Incidentally, [31] proposed taxonomies for both questions and answers. Fundamentally, their question taxonomy extended [10] by adding a social category, which comprises queries that seek interactions with people. They discovered a high correlation between answer and question types. More specifically, constant questions are more likely to target factual unique answers, while opinions get subjective answers.

In this paper, we consider the relationship between search logs and Yahoo! Answers pages connected via user clicks as additional source for the session boundary analysis, cf. sec. 3.

2.4. Semantic Categorization in cQA

Our study focuses on the semantic categorization of question-like search queries, which cover a wide variety of informational queries that do not necessarily bear named entities. In particular, this paper studies the impact of preceding queries in user sessions for tackling the lack of context in this semantic categorization. Our approach is supervised, trained with a large set of automatically tagged samples via inspecting click patterns between search queries and Yahoo! answers questions.

In [32] we proposed a novel category-specific learning to rank approach for effectively ranking paraphrases for cQA, and empirically demonstrated that the question categories dramatically affect the recall and ranking of past answers. For example, it is possible to use the user generated questions of a cQA along with search engine query logs to automatically formulate effective questions or rank paraphrases in order to improve search in cQA [1, 29]. A major advantage of such a query-to-question expansion approach for cQA is that it can help to retrieve and order more related results from cQA archives and hence, can improve the search accuracy. We obtained empirical evidence that the subjective and objective nature of cQA questions substantially impacts on the detection of paraphrases that are effective in boosting the recall and ranking of past answers, due to the strong connection between categories and both question intents. Our results unveiled that retrieval and ranking of social media data can be improved when category information is used.

In this previous approach, we assumed the categories for all questions are given by the users, which actually is the case for many cQA services like Yahoo! Answers, where a questioner has to select and add a category to her question from a prespecified list of categories. Hence, the focus of the research was on the exploitation of category

No.	Search query	Clicked hosts	Search query	Clicked hosts
1	you tube how do i make a heel strap	**Beauty & Style**	0.10n hno2	
2	cracked heel repair		0.10n hno2 acidic?	
3	wraps for cracked heel repair	pantryspa.com	0.10n hno2 basic	www.jiskha.com
4	oil based moisturizer brands		0.10n hno2 neutral?	
5	oil based moisturizer cream brands	ezinearticles.com	0.10n hno2 acidic	
6	oil based moisturizer cream brands	www.alibaba.com	hno2 acidic	wiki.answers.com
7	oil based moisturizer heel cream brands	www.amazon.com	nano2 acidic	
8	oil based moisturizer heel cream brands		nano2 acidic or basic	wiki.answers.com
9	oil based heel cream		nano3 acidic or basic	**Science & Maths**
10	is vaseline considered a oil based moisturizer	**Beauty & Style**	nh4no2 acidic or basic	chemicalforums.com
				www.jiskha.com
				Science & Maths
11	vaseline uses	www.ehow.com	nano3 acidic or basic	www.chacha.com
				www.legacy.com
				Science & Maths
12	is vaseline an oil moisturizer	**Beauty & Style**	nano2 acidic or basic	**Science & Maths**
13	goodle	www.google.com		

Table 1: Two samples of transactions: one comprises 13 queries and the other 12 (categories are shown for clicked Yahoo! Answers pages).

information for ranking paraphrases, not on automatically assigning a semantic category to an unclassified search query. However, many of these cQA services also provide a standard information retrieval API, which helps users to search in the cQA archive by formulating question-like keyword-based search queries. However, these search queries are unclassified, and thus semantically unspecified. Consequently, a natural further research question is whether it is possible to automatically predict the categories of new question-like search queries and how effectively this can be achieved by means of a Machine Learning approach that would be able to make use of a cQA archive in a fully automatic manner, i.e., without the help of manually inspected, cleaned and optimized data.

3. Corpus Acquisition

In order to carry out our study, we automatically built a corpus by means of integrating Yahoo! Search query logs with Yahoo! Answers pages. This integration is on the basis of click patterns across user search sessions. In particular, we focused our study on search queries in English submitted in the United States from May 2011 to March 2013. We assume that user clicks to Yahoo! Answer pages signal that, at some point during these search sessions, users prompted questions and discovered pertinent information on the visited Yahoo! Answer pages. Note that search engines provide the first lines of best answers as snippets in the respective result pages, when hits come from Yahoo! Answers. Overall, we extracted about 71 millions full user sessions containing questions by keeping only those elements connected to Yahoo! Answers via at least one user click.

h	Search query	Clicked host(s)	Search query	Clicked host(s)
10	10 dpo and rib pain		1980 chips	
9	rib pain could i be pregnant		1980 chips snack	chowhound.chow.com
8	rib pain 2 weeks pregnant		conquistos	**Food & Drink**
7	do ribs hurt after ovulation	**Pregnancy & Parent.**	conquistos	
6	rib pain after ovulation		conquistos chips	
5	rib pain before positive preg test	pregnancyforum.co.uk **Pregnancy & Parent.**	conquistos chips	
4	will my uterus expanding make my ribs hurt		who made conquistos chips	
3	2 weeks pregnant and ribs hurt		who made conquistos picante chips	
2	rib and back pain after implantation		who made conquistos picante chips	
1	is rib pain a sign of pregnancy		conquistos picante chips	www.inthe80s.com
0	is right rib pain a sign of pregnancy	**Pregnancy & Parent.**	corn quistos picante chips	**Food & Drink**
10	1950's prices list for soda		0.25 simpilist form	
9	1950's minimum wage	www.sodahead.com	0.25 simplest form	www.chacha.com
8	1950's minimum wage	wiki.answers.com	how do i write 2.50 in simpilest	
7	1950's minimum wage		how do i write 2.50 in simplest	www.ehow.com
6	utility cost in 1950	wiki.answers.com	how do i write 2.50 in simpilest	
5	utility cost in 1950	**Business & Finance**	how do i write 2.50 in simplest	**Education & Refer.**
4	utility cost in 1950	wiki.answers.com	what is the fraction for 2.50	wiki.answers.com
3	utility cost in 1950		3.4 written fraction form	
2	price of burger and milkshake in 1950	aolanswers.com	4 3/10 written in fractions	
1	price of burger and milkshake in 1950	aolanswers.com	4 3/10 written in decimals	www.blurtit.com
0	price of burger and milkshake in 1950	**Dining Out**	how do i write 5/20 as a decimal	**Education & Refer.**
10	0.05 m solution of substance		1200 diabetic diet plan	
9	substance with a density greater than one house f hold		diabetic diet plan	
8	substance with a density greater than one house hold		diabetes and choclolate	www.sugarstand.com abcnews.go.com
7	substance with a density greater than one house hold	Science & Maths	diabetes and alcohol consumption	diabetes.webmd.com www.livestrong.com diabetesjournals.org
6	iron oxide household		hypo thyroid and symptoms of diabetes	rightdiagnosis.com www.lifescript.com
5	iron oxide household name	wiki.answers.com	hypo thyroid and medication	www.webmd.com
4	homozygous substance with exactly 29 protons	wiki.answers.com	hypo thyroid and thyroid 1 g medication	
3	commom inert gas house hold	wiki.answers.com www.chacha.com **Science & Maths**	hypo thyroid and thyroid 1g medication	
2	commom inert gas household		thyroid 1g medication	
1	common inert gas household		amour thyroid 1g medication	www.rxlist.com www.medhelp.org **Health**
0	common inert gas household	wiki.answers.com **Science & Maths**	what is amour thyroid 1g?	**Health**

Table 2: Six eleven-queries transactions corresponding to six distinct Yahoo! Answers categories. In the case of clicked Yahoo! Answers pages, categories are shown instead of hosts.

However, search sessions can cover a large period of time, thus they can comprise a wide variety of search needs, and as a logical consequence they can contain a large number of queries. For this reason, we looked at smaller units (also called transactions) in these sessions that are likely to aim at one goal.

We split each search session into transactions by means of two criteria. First, we benefited from the time difference by which two consecutive queries were sent to the search engine. We used a gap of 300 seconds as a transaction splitter, assuming that longer periods of time indicate that users are likely to have changed their search needs. It is worth emphasizing that the size for this temporal cut-off has been popularly used for segmenting query logs cf. [18].

Secondly, conventionally, navigational queries (e.g. , "*twitter*") are prompted by users when they want to reach a particular web-site they bear in mind. As a rule of thumb, most frequent queries in search logs are navigational [10, 27]. Therefore we used all search queries having a frequency higher than 1,000 across our session corpus as

additional transaction splitters.

Next, in order to study the impact of preceding queries in the session on the semantic tagging of a new submitted question-like search query, we kept only transactions containing at least eleven queries, where a user click links the eleventh or a later query with Yahoo! Answers, and hence with one of its categories. In other words, we studied the impact of up to ten historical queries.

Table 1 shows two transactions, one consisting of 13 and the other of 12 queries. Each line in this table consists of a query index (from 1 to 13), and the search string as well as its clicked hosts (which can be empty, a single click or a sequence of clicks). Note that several eleven-element transactions can be derived from the rightmost transaction. In this sample, two query sequences: 1-11 and 2-12 are acquired, since the queries with number eleven and twelve are connected to Yahoo! Answers. In juxtaposition, solely one sequence is harvested from the other transaction: 2-12. Overall, we obtained 783,528 smaller transactions containing only eleven elements, in which the 11th query is related to Yahoo! Answers by means of a user click. Table 2 displays six samples belonging to six different Yahoo! Answers categories. For each of these instances, we listed their corresponding clicked hosts.

Table 3 shows the distribution of Yahoo! Answers top-level categories across our acquired corpus. The three most prominent semantic types were Science & Maths (27.83%), Health (8.83%) and Educacion & Reference (8.26%), while the less frequent categories are News & Events (0.18%), Environment (0.18%) and Dining Out (0.17%).

Category	%	Category	%	Category	%	Category	%
Science & Mathematics	27.83	Entertainment & Music	3.86	Social Science	2.35	Yahoo! Products	0.65
Health	8.83	Arts & Humanities	3.68	Home & Garden	2.10	Local Businesses	0.19
Education & Reference	8.26	Pregnancy & Parenting	3.61	Food & Drink	2.06	News & Events	0.18
Business & Finance	5.37	Cars & Transportation	3.15	Travel	1.99	Environment	0.18
Society & Culture	4.07	Beauty & Style	2.94	Games & Recreation	1.70	Dining Out	0.17
Family & Relationships	4.05	Pets	2.76	Sports	1.69		
Politics & Government	3.99	Computers & Internet	2.71	Consumer Electronics	1.64		

Table 3: The distribution of Semantic Categories across our Acquired Corpus (%).

4. Features

In our work, we study the effectiveness of several kinds of properties derived from the target search query (the element being classified) and from its previous historical elements in the transaction. In this paper, the target query can also be referred to as the eleventh query in its transaction, the last query or h_0 (see table 2). For the sake of clarity, from now on, a transaction only refers to a sequence of eleven queries in our corpus. In our models, we make use of eleven different types or groups of features covering surface, lexical, linguistic and semantic characteristics of the corpus, cf. table 4. They will now be introduced and described in the next paragraphs.

Type	Brief Description
Bag-of-words (BoW)	Different alternatives: raw terms, lemmata, spell correction, with and without stop-words.
Latent Topic Models	Top three topics determined by LDA and PLSA.
Semantic Analysis	ESA, twenty-eight WordNet types of relations, four kinds of synonyms (nouns, adjectives, verbs and adverbs), and eight sorts of collocations.
Lexical Chains	Modelling one term: adjective, nouns, verb, adverb; and different terms: collocations and semantic relations.
Acronyms	Two resources were used for acronym resolution: acronymlist and allacronyms. The latter was additionally exploited for obtaining indicators such as categories and tags.
String analysis	The number of unique queries, the highest frequent query, if the target query embodies or is embodied in another query. We also check streak of queries or if it was previously asked. We also count words with and without stop-words.
String distances	We benefited from twelve distance metrics between the target query and its preceding items.
NLP	We identified 26 named entity types and the most frequent class. Counts related to 46 POS categories were considered.
Yago2s	Abtractions provided by the labels contained in the rdfs:subclass hierarchy.
Wikipedia	We profited from structures provided by aliases, sense indicators, categories, infoboxes pairs attribute-values, and WEX FreeBase mappings.
Yahoo! Categories	Semantic categories of previously clicked Y! Answer pages.

Table 4: Bird's-eye view of the characteristics tried in our models.

Bag of Words. The first array of features takes into account different variations of the bag-of-words (BoW) approach. In the first place, we considered its version with the unmodified tokenized terms. But also, we accounted for several alternatives that capitalizes on linguistic processing such as lemmatization and spell correction [33]. For spell correction, we benefited from Jazzy[4], whereas for lemmatization and tokenization we used Montylingua[5]. When performing spell correction, we picked the most frequent alternative, proposed by Jazzy, that appeared within the previous or the target query. In addition, we also took into account variations with and without stop-words. Note that the respective BoW(h) is constructed on top of a sequence of h preceding queries. Here $h = 0$ means that it solely comprised the target query, while $h = 5$ means that this BoW was merged with the word frequencies of the five previous queries. From now on, h is used for denoting the amount of preceding queries.

Latent Topic Models. The second group of features is extracted from latent topic models. More specifically, we profited from two different strategies: Latent Dirichlet Allocation (LDA) [34] and Probabilistic Latent Semantic Analysis (PLSA) [35]. For the former, we took advantage of the implementation by GibbsLDA[6] and for the latter, we capitalized on a publicily available implementation in Python[7]. We consider features of the form lda(h,p) and plsa(h,p) corresponding to the assignment at the p-th position outputted by each model, respectively. In these attributes, we account solely for the first, second and third ranked elements ($p = 1, 2, 3$). The value of h signals the number of

[4] www.spellcheck.net/jazzy

[5] web.media.mit.edu/ hugo/montylingua

[6] gibbslda.sourceforge.net

[7] http://www.mblondel.org/journal/2010/06/13/lsa-and-plsa-in-python/

preceding queries utilized for modelling the latent topics. That is to say, $h = 10$ means that we profited from all available context in the transactions, whereas a value of zero indicates that we only use the query being classified. For example, in the attribute-value pair <lda(4,2); 15>, the 15-th latent topic ranked second, when target queries and their respective previous four historical elements were amalgamated for modelling LDA topics. Note that all these models were constructed on top of BoWs without stop-words and modelling twenty-six latent topics, which is the number of target question categories.

Semantic Analysis. The third set of attributes is harvested from conducting semantic analysis. For this purpose, we benefited from three distinct types of resources: Explicit Semantic Analysis[8] (ESA), WordNet[9] and the Oxford Collocation Dictionary[10]. The properties distilled from each of these three sources are as follows:

- Note that ESA represents the meaning of any text as a weighted vector composed of the top-k related Wikipedia-based concepts [36]. This semantic space has shown to be particularly useful for coping with some short documents [37]. Similarly to previous features, we devised an attribute, esa(h,k), which performs this analysis considering the last h queries ($h = 0, \ldots, 10$) and adds the k top scored Wikipedia concepts to the feature vector. We tested models accounting for different numbers of related concepts ($k = 1, \ldots, 10$).

- Additionally, we exploited twenty-eight different sorts of semantic relations provided by WordNet. More precisely, we checked if a semantic connection between a pair of terms exists. WordNet contemplates types such as hypernyms (e.g., pressure→distress), holymns (e.g., professor→staff), hyponyms (e.g., pressure→oil/gas pressure) and meronyms (e.g., service→supplication). Accordingly, we added a boolean attribute, wordnet(h,X), signalling whether or not the target query was expanded with terms linked by the relation "X", whenever this relation holds between a term in the target query and in any other query within the h previous queries. Along the same lines, we also took advantage of WordNet for finding putative synonyms within transactions. Each pair of synonyms found by WordNet was validated via imposing that their their part-of-speech (POS) tag categories must coincide. We considered the following classes of synonym pairs: adjectives, nouns, verbs and adverbs. In so doing, we mapped Penn TreeBank tags returned by Stanford caseless models[11] to their respective universal POS categories [38].

- As for collocations from the Oxford Dictionary, analogously to WordNet relations, we accounted for a property, collocation(h,X), indicating if the query was expanded with the words connected by the type of collocation "X". Specifically, this dictionary yields eight kinds of relations: adjectives (e.g., revival→great), adverbs (e.g., pack→carefully), following and preceding verbs (e.g., outbreak→lead and outbreak→occur), prepositions (e.g.,

[8] ticcky.github.io/esalib/

[9] wordnet.princeton.edu

[10] oxforddictionary.so8848.com

[11] nlp.stanford.edu/software/tagger.shtml

revival→of), quantifiers (e.g., revelation→flash), and related nouns (e.g., rib→cage) as well as verbs (e.g., rib→crush).

Lexical Chains. The notion of lexical chains inspired the fourth battery of features. In short, a lexical chain is a sequence of related terms, typically spanning adjacent words or sentences, that have a correspondence to the structure of the text, and aid in resolving ambiguity as well as assist in interpreting the underlying concept represented by a term [39, 40]. In this study, we perceived as lexical chains sequences of related terms within sessions that show some syntactic patterns. We computed these chains by tracking the position in which a term shows up in the transaction. We distinguished three different positions within search queries: begin (denoted as B), end (E) or others (O). We consider two occurrences of the same term as member of the same chain if and only if their POS categories match. More precisely, their syntactic classes were mapped to their respective universal tags, whereby chains for terms corresponding to four syntactic roles were built: verb, noun, adjective and adverb. A good example is the characteristic lex-chain-noun(chips_O_E,9) (see table 2), by which we model a chain composed of the term "chips" appearing as noun in the end(E) and elsewhere(O) across a transaction of size h=9. In the same vein, in table 2, we find lex-chain-noun(household_O_E,5) and lex-chain-noun(thyroid_B_O,2). Accordingly, we used four boolean attributes to indicate whether or not the feature vector was enriched with the chains discovered for the respective universal syntactic category.

For the case of related terms, we did not only consider the same element bearing the same syntactic category, but also we contemplated lexical chains of distinct terms, but bearing a semantic or collocation relation triggered by WordNet and the Oxford Dictionary, respectively. For example, lex-chain-col-adj(inert+gas_O,0) indicates that the collocation of type adjective observed by the noun "gas" and the adjective "inert" was discovered in the target query. In like manner, for each of these types of semantic/collocation relations, we added a boolean property denoting whether or not the respective chains were incorporated into the feature vector. Since these new relations can bear several different terms, we modelled this diversity by means of an attribute representing the amount of distinct elements in the respective chain. For example, the previous chain is composed of one adjective and one noun, with both words appearing "elsewhere" within the query, thus lex-chain-col-adj-value(inert+gas_O,10) is equal to two. Note that we use lemmata of query terms for the look-up of their entries in these resources.

Acronyms. The next group of properties is derived via resolving acronyms. Sometimes, users submit queries embodying both phrases and their respective acronyms during their search sessions in order to achieve their goals [26]. Since this relation can give useful hints for semantically categorizing the target query, we capitalized on two acronym resolution on-line databases: www.allacronyms.com and www.acronymlist.com. By crawling both web-sites, we obtained 849,473 and 44,369 abbreviation-resolution pairs, respectively. In the case of www.allacronyms.com, resolutions can be associated with eleven categories (e.g., technology, military and education) and 1,906 tags (e.g., airline, school, journal, and computer). Consequently, we took into account all this information to produce the following characteristics:

- A binary attribute allacronyms-res(X,h) indicating whether or not the resolution of a term X was found within the h previous queries in the session. Similarly, we use a property acronymlist-res(X,h) when the matching was provided by www.acronymlist.com instead of www.allacronyms.com. Here, X can be any term in the target query, except from stop-words.

- If the previous feature is true, this means the resolution of a term X was discovered, then two extra binary attributes are taken into account: allacronyms-cat(Y,h) and allacronyms-tag(Z,h). The former denotes the occurrence of the category Y provided by the identified relation, while the latter in the event of the tag Z.

- Even though the resolution of X fails, we make allowances for a boolean property allacronyms-related-tag(Z,h), which signals if the tag Z, related to target query term X, is contained in any of the h previous queries. Note that we expect this relation to be weaker, since the putative acronym X was not resolved during the sequence of h queries.

String Analysis. Further, we examined the effectiveness of a set of string-based features, mainly encouraged by the findings of [22]. In particular, they observed that half of consecutive query pairs are identical. In their study, they also noticed that the number of discovered patterns producing longer reformulations double the amount for shorter reformulations. We modelled this user behaviour as a means of investigating if a connection to the semantic category of the query exists. In effect, we designed the following attributes:

- unique-queries(h) counts the number of unique queries during the last h submissions. In the same vein, highest-frequency(h) denotes the frequency count of the most recurrent query in the transaction (one in the event that all queries are different). Take for instance the third sample in table 2, we obtain unique-queries(5)=2 and highest-frequency(6)=4.

- is-subquery(h) and has-embodied(h) are the number of queries that contain and are contained in the current query, respectively.

- contains and is-contained are two boolean features indicating if the target query embodies or is embodied in its previous query (h=1), respectively. We also inspected if the target query is identical to its prior query, but it just inserts or deletes one word (represented by the features has-inserted-token and has-deleted-token). An illustrative case is depicted in the first sample in table 2, where we obtain has-inserted-token=true.

- streak-queries(h) checks if the target query belongs to a streak of identical queries. More precisely, it counts the number of queries involved in this streak. Along the same lines, asked-before(h) signals if the target query has been prompted anywhere across the h previous queries. In the fifth sample of table 2, we have streak-queries(2)=2 and asked-before(10)=true.

- In addition, we took advantage of several word frequency statistics, such as the number of tokens and unique terms in the transaction. Here, counts with and without stop-words were considered. Furthermore, we calculated

these statistics at the level of target query level only and also in conjunction with the prior h queries. To illustrate, the second sample contains number-tokens(4)=21 and unique-terms(4)=8, whereas their counterparts without stop-words are number-tokens-wsp(4)=18 and unique-terms-wsp(4)=7, respectively.

String Distance. In the same spirit of [24], we benefited from twelve string distance metrics implemented by the Second String[12] package (e.g., Levenstein, Monge-Elkan, Jaro-winkler and Jaccard). These metrics were utilized for computing the distance between the target query and each of its h predecessors. Thus, twelve boolean attributes denote whether or not distances provided by the respective metric were incorporated into the feature vector.

NLP-based Features. Moreover, we employed two Natural Language Processing tools to extract attributes: a named entity recognizer specialized in search queries (NERQ), and POS tagger. As for NERQ, we made use of the two-step technique adopted by [41], while the Stanford caseless model[13] for obtaining POS tags. This NERQ tool is capable of distinguishing twenty-six distinct classes of named entities within search queries like beverage, brand name, business, company ticker, cooking method, cuisine and many others. This cooperated on adding one property per entity class (e.g., nerq-food(h)) indicating the amount of entities of each kind within the query span given by h. Further, we also used the named entities found for expanding the target query, that is to say, nerq-names(h) is a boolean property that signals whether or not the target query was expanded with the named entities discovered across the query span framed by h. Furthermore, we considered a feature that adds to our model the class of the highest frequent entity type in the span confined to h. As for POS, we added one feature per Penn Treebank word category indicating the number of tokens bearing the respective class within the query span enclosed by h. To illustrate, the feature pos-NNP(3)=5 denotes that five tokens categorized as proper nouns within the span comprising the target and its preceding three queries.

Yago2s. The next battery of attributes distils from the Yago2s[14] database, in particular from the linked data contained therein. In recent years, there have been important advances in exploiting linked data for semantic search and question answering [42]. In order to take advantage of this resource, we extract unigrams, bigrams, trigrams and tetragrams from queries, whereby we look-up for rdfs:label predicates in this semantic database. We single out n-grams that are likely to be words and named entities by checking if they are linked in Yago2s to a WordNet domain or not. In the case of words, we navigate through the predicates corresponding to the rdfs:subclass hierarchy until the top node. While walking up in the hierarchy, we validate each level by checking if any of its predicates is supported by the span enclosed by the h queries. In other words, a level is supported if any of its objects appears in the sequence of queries. In the event of entities, we proceed is a similar fashion, but considering the categories linked by the rdfs:type predicate. To exemplify, the yago2s(l,h)=V denotes that the object V is supported by the sequence of h queries, and this object is up l levels in the hierarchy from a term X in the target query.

[12] secondstring.sourceforge.net

[13] nlp.stanford.edu/software/tagger.shtml

[14] www.mpi-inf.mpg.de/yago-naga/yago

Wikipedia. Also, we use the aforementioned look-up strategy to discover related-entities across Wikipedia so that we can profit from their respective aliases, sense indicators, categories and infoboxes for devising the following features:

- alias(id,*h*) means that the Wikipedia page id was found to be referred by an entity in the target query and an alias within the span confined to the sequence of the *h* queries. Take for instance, the query terms 10,000 and 10000 are connected via their common reference to the Wikipage number 19686545. We made use of a boolean feature denoting whether or not the target query was expanded with the discovered reference numbers.

- We refer as to sense indicators those keywords usually added in parentheses to the titles of Wikipedia pages. These keywords play the role of helping to disambiguate the main concept. To illustrate, the titles "*James Clark (artist)*" and "*James Clark (programmer)*" were added with the sense indicators artist and programmer, respectively. This way it becomes clear that they refer to people from different domains of expertise. Like aliases, we exploit a boolean property signalling whether or not the target query was expanded with sense indicators found in respective sequence of preceding queries.

- By the same token, Wikipedia pages are coupled with a list of categories, which cooperate on establishing relation amongst documents according to their topics and on fetching related information. Therefore, a boolean attribute modelled whether or not the target query was expanded with categories supported by respective sequence of queries.

- As for infoboxes, we designed two attributes. The first one works in a similar way to the sense indicators, but this time we capitalized on infobox types. The second is also a boolean value indicating whether or not the current query was expanded with the values supported by the respective span of *h* queries. These values corresponds to the attribute-value pairs provided by the retrieved infoboxes.

We extended the use of Wikipedia by capitalizing on WEX[15], which provides mappings between Freebase topics and Wikipedia articles. Since this Freebase taxonomy has two-levels, one feature signals the use of its first level, whereas another the second. For instance, the query "*waubonsie valley high school*" is linked with the FreeBase first level categories: education, location and organization; while at the same time, with full classes like education→educational_institution and education→school. Note that both features are boolean, meaning that the target query was expanded with these classes, whenever these were supported by the respective sequence of *h* queries.

Yahoo! Answers Categories. Lastly, we took advantage of previously categorized queries within search sessions. In some occasions, users are targeting at some particular topic, and hence they might view several pages from Yahoo! Answers in a row, which might be semantically related questions in the cQA archive. In order to model this behaviour, we added boolean attributes indicating if any of the twenty-six top-level Yahoo! Answers categories appears across

[15] wiki.freebase.com/wiki/WEX

the Yahoo! Answers links clicked by the user in the corresponding span of h queries. In all samples depicted in table 2, we find a preceding categorized search query within the transaction. Note that categories differ in one case: Dinning Out and Business & Finance.

5. Experiments

As for learning models, we used Maximum Entropy (MaxEnt)[16] classifiers with Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) and L_2-regularization. This is a popular combination for parameter estimation, when dealing with a large number of features, especially for fitting log-linear (MaxEnt) models[43, 44]. As secondary learning models, we capitalized on Winnow2 classifiers[17]. Basically this sort of online learner updates its linear discriminant function, whenever a training example is incorrectly labelled, making it a computationally efficient algorithm in both time and space [45]. In all our experiments, we carried out a three-fold cross-validation operating on the same three equally-sized random splits.

Since both learning models output a confidence value for each candidate label, in our case for each of the twenty-six target question categories, we took advantage of the *Mean Reciprocal Rank* (MRR) for assessing the performance of our models. Basically, this metric is the multiplicative inverse of the position in the confidence ranking ($rank_i$) of the first correct label [46]. The MRR is then the average of the reciprocal ranks of the predictions obtained for a sample of queries (Q):

$$\frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}.$$

A **greedy** algorithm (a.k.a. Sequential Forward Selection or SBS) was used for selecting the best array of features for each learner [47]. This process starts with an empty bag of properties and after each iteration adds the one that performs the best. In order to determine this feature, this procedure tests each non-selected attribute together with all the properties in the bag. The algorithm stops when no non-selected feature enhances the performance. In essence, our work studies the impact of a large number of fine-grained attributes, which on the one hand, it helps to analyze each particular contribution to the task, and ergo to draw better conclusions; but on the other hand, it makes feature selection computationally demanding. As a matter of fact, the amount of properties significantly grows when exploiting large spans of previous queries (higher h values), because each feature is computed considering all plausible levels of history confined to the respective span. To illustrate this, think on the raw BoW(h) attribute in conjunction with the span of queries given by $h = 3$. This combination produces four candidates properties: a) BoW(0) which solely considers the words contained in the target query; b) BoW(1) the words within the target query with its predecessor; c) BoW(2) the words in the target query together with its two previous queries; and analogously d) BoW(3) with the three predecessors. The reason to opt for this approach is assuming that we do not know, for our classification task,

[16] www.logos.t.u-tokyo.ac.jp/~tsuruoka/maxent/

[17] We ported the multi-class implementation provided by MALLET to C++ as a means of gaining computational efficiency.

h	0	1	2	3	4	5	6	7	8	9	10
MaxEnt	0.8441	0.8443	0.8451	0.8456	0.8467	0.8473	0.8480	0.8486	0.8493	0.8493	0.8510
Winnow2	0.7487	0.7557	0.7614	0.7640	0.7675	0.7712	0.7729	0.7684	0.7742	0.7751	0.7993

Table 5: Results accomplished by the best feature determined by SBS. Each cell indicates a different combination of a learning model and a level of context. All numbers correspond to MRR scores.

how much context/history is good for getting the best out of each feature. In actuality, in our illustrative example, it might be that adding queries is detrimental to the performance from some certain point on. In brief, this large amount of attributes shapes a vast search space to be tested by the greedy algorithm. For this reason, all features that decreased the performance were removed after the fourth iteration. Note that, from this point, properties added to the bag of selected characteristics bring relatively smaller gains. In practical terms, these attributes bring about an increase lower than 0.05%. Hence, starting to removing attributes at this point lessens the possibility of missing a highly discriminative attribute, while at the same time, considerably reducing the size of the search space, especially taking into account that training times turn to be much larger as long as more candidate attributes are already in the bag.

Since we want to quantify the impact of the previous queries on the semantic classification rate, we built a **baseline** by taking into account target queries only. We run the greedy selection algorithm so as to determine the best performing model. Thus, using previous queries worth the effort if and only if we can do better than accounting solely for target queries. Overall, table 5 displays the improvements reaped by both learners when accounting for different levels context.

In substance, our empirical results point out to the following findings:

1. Both models finished with the highest MRR score by means of exploiting all ten preceding queries. Particularly, in the event of MaxEnt, this increase reached 0.8%, whereas 6.76% in the case of Winnow2. Therefore, corroborating the positive contribution of the contextual evidence harvested from prior queries.
2. In their essence, both MaxEnt and Winnow2 are specialized in dealing with large feature sets. In our task, MaxEnt models outclass Winnow2 regardless the number of previous queries considered when building the models. Notably, this difference in performance decreases as long as Winnow2 models are constructed on the basis of a larger context. Specifically, this gap in performance is reduced from 12.74% to 6.47%, when enlarging the context from zero to all prior ten queries. It is worth emphasizing here that Winnow2 is more efficient in both training and classification times.
3. Except from one case, increasing the context on a query by query basis brought about a growth in performance. Ergo, it can be stated that, as a rule of thumb, the classification rate goes up in tandem with the amount of preceding queries yielded as context. That is to say, it is possible to mine fine-grained linguistically-oriented semantic features for modelling this additional contextual evidence properly.

Category	↑ (%)	↓ (%)	Δ (%)	Category	↑ (%)	↓ (%)	Δ (%)
Science & Maths	3.38	2.85	0.25	Computers & Internet	10.67	7.66	1.91
Health	8.13	6.55	0.76	Social Science	17.21	13.51	2.00
Education & Reference	11.96	10.56	0.54	Home & Garden	12.29	9.16	1.72
Business & Finance	9.40	7.19	1.16	Food & Drink	11.47	8.74	1.42
Politics & Government	13.38	10.88	1.40	Travel	16.33	13.87	1.17
Family & Relationships	10.07	9.10	0.40	Games & Recreation	15.45	10.72	2.95
Society & Culture	16.07	14.19	1.22	Sports	15.49	10.38	3.31
Entertainment & Music	14.02	10.71	1.86	Consumer Electronics	11.38	8.31	1.80
Arts & Humanities	16.66	13.90	1.37	Yahoo! Products	18.79	15.38	0.82
Pregnancy & Parenting	10.50	8.88	0.71	Dining Out	21.78	20.25	0.95
Cars & Transportation	7.88	5.23	1.53	Local Businesses	33.69	30.68	-1.21
Beauty & Style	11.42	8.72	1.38	News & Events	30.32	26.91	2.24
Pets	6.88	4.38	1.25	Environment	27.82	22.43	2.24

Table 6: Gains and losses per category obtained by the best model over the baseline (MaxEnt). Percentages of improved (↑) and worsened (↓) cases. Also, increases/drops are signalled in terms of percentage variations of MRR score (Δ).

4. Remarkably, adding the tenth historical query resulted in a substantial gain for both learners with respect to their prior configuration. To be more exact, this addition enhanced the MRR score by 3.1% for Winnow2. Although, this betterment is much smaller for MaxEnt (0.2%), it is still substantially larger with respect to the improvements achieved by exploiting prior levels of context. In light of this, we deem that higher classification rates can be achieved if our models are enriched with more contextual information, e.g., the content of previously visited cQA or web pages.

In summary, our findings point out to the the positive impact on the classification rate of considering large number of preceding queries as contextual evidence, especially for extracting effective attributes that boost the performance of cost-efficient models such as Winnow2. Further, our outcomes highlight the fact that MaxEnt is a more suitable, but at the same time more computationally demanding, learner for this task.

5.1. Analysis Per Category

On a different note, table 6 underlines the results from the viewpoint of each category. More precisely, the improvement or diminishment accomplished by MaxEnt ($h = 10$) over the baseline MaxEnt ($h = 0$) per category. This table shows three columns per category, denoting a different percentage variation with respect to the baseline: improved (↑) and worsened (↓) instances as well as the increases/drops in terms of MRR score (Δ). These figures reveal the following insights:

1. For all but one case, making allowances for context provided by the session enhanced the MRR score. The exception regards the category "*Local Businesses*". In this case, the classification rate worsened 1.21%. Con-

versely, the categories "*Sports*" and "*Games & Recreation*" reaped the highest growths, 3.31% and 2.95%, respectively.

2. Together with the overall decrease for "*Local Businesses*", we also see that the performance improved and worsened for 33.69% and 30.68% of its queries, respectively. From these drops, the categories "*Businesses & Finance*" (21.40%) and "*Travel*" (19.87%) take the bigger shares. Interestingly enough, the improvements came from queries that were tagged as "*Businesses & Finance*" (23.86%) and "*Travel*" (17.10%) by the baseline, and now perceived as "*Local Businesses*" by the best MaxEnt model. In light of this, we can conclude that the semantic range of these three categories overlaps to a substantial degree, making much more difficult and uncertain to distinguish the right category. Note that these three categories focus their attention on businesses and locations, but with different goals.
3. For all categories, the number of instances that improved the performance surpassed the amount of cases it dropped. To illustrate, 8.13% of the instances embedded in "*Health*" obtained a better ranking, whereas for 6.55% of the samples the performance was detrimental. These betterments were due to samples that the baseline conceived as "*Science & Maths*" (30.52%), "*Pregnancy & Parenting*" (16.35%) and "*Beauty & Style*" (11.85%). Take for instance the following examples:
   (a) The baseline interpreted the query "*abdominal pain but delayed period*" as pertaining to the category "*Pregnancy & Parenting*". Conversely, the best MaxEnt model had access to hints conveyed in the previous queries such as "*white blood cells high*" and "*uti infection*", which assisted in discriminating its right label.
   (b) In the same vein, the query "*leaving toenail polish on too long*" was labelled as "*Beauty & Style*" by the baseline, but cues yielded in preceding queries including "*home remedy*" and "*toenail ringworm*" helped the best MaxEnt model to select the "*Health*" class.
4. Noteworthily, the categories "*Sports*" and "*Games & Recreation*" finished with the highest MRR growths, 3.31% and 2.95%, respectively. Specifically, the best MaxEnt model correctly classified 5.11% (15.49%-10.38%) more instances than the baseline, when coping with "*Sports*", while 4.73% in the event of "*Games & Recreation*". For the latter, these improvements were due to a higher recognition rate with respect to categories such as "*Consumer Electronics*" (10.47%) and "*Computers and Internet*" (10.32%). For the former, we find "*Health*" (11.77%) and "*Entertainment and Music*" (9.48%). To exemplify this, take the next cases:
   (a) The baseline associates the query "*health benefits of boxing*" with the category "*Health*". However, when the session is analysed, semantic cues such as "*sports*", "*equipment*" and "*cycling*" show up in the context, cooperating on distinguishing the right label (*Sports*).
   (b) Along the same lines, the query "*connecting xbox to mac with router*" was perceived as "*Computers & Internet*" due to concepts including "*connecting*" and "*router*". In the session, we discover semantic trails like "*saving*" and "*dlc*", which aid in allocating this query in the right class "*Games & Recreation*".

In conclusion, the analysis of our results show that linguistically-oriented semantic cues, harvested from sessions, are useful for ameliorating the semantic classification of question-like search queries. Our figures emphasize their effectiveness at bettering the ranking and at increasing the ratio of the number of improved to worsened cases. On the whole, our analysis underscores that profiting from the session helps to reduce the semantic range of otherwise ambiguous queries. In addition, it underlines the usage of contextual evidence as a contributor to enhance the performance regardless of the semantic category, in general.

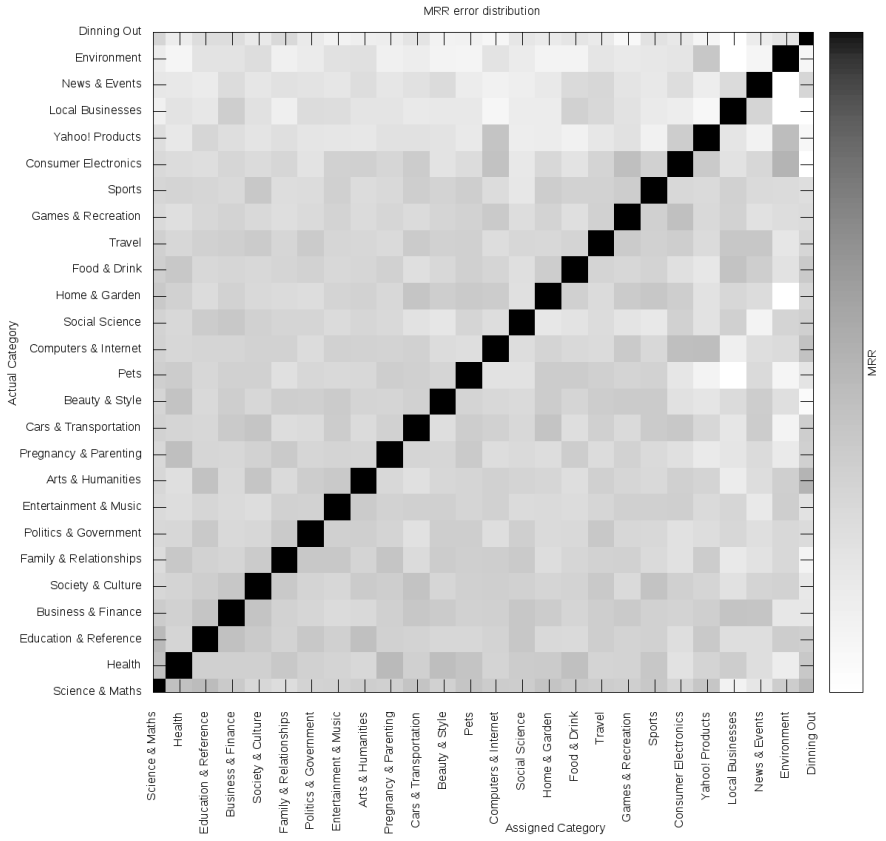

Figure 1: Heat map displaying the MRR scores, achieved by the best MaxEnt model, for each pair of actual and assigned label. The black diagonal denotes correct classifications.

Incidentally, figure 1 shows a different angle of the performance per category. This heat map displays the MRR score achieved by the best MaxEnt model with respect to pairs of actual and assigned labels. In other words, it compares the MRR score achieved by each actual label in conformity to a particular assigned class. Put differently, the average reciprocal rank obtained by a label X when the assigned label was Y. For instance, a cell (actual=X,assigned=Y)=0.5 means that the actual label X always ranked second when the assigned label was Y. For

Feature Group	Winnow2	MaxEnt
Bag-of-words	without-stopwords(11), lemmatized-without-stop-words(2), lemmatized-spell-correction-without-stop-words(2)	without-stopwords(11), lemmatized(11)
Semantic Analysis	ESA(11)	ESA(11) WordNet: hyponyns(11), hypermyns(6), word-forms(5) Collocations: adjective(1), preceeding-verbs(11), preposition(11), related-nouns(2)
Lexical Chains	collocation-adverb-value(1)	noun-chains(2), wordnet-ttributes(1)
Acronyms	acronymlist-category(1)	
NLP	POS-Taggings: CC(1), LS(1), UH(1), ”(1)	POS-Taggings: CD(1), NNP(2), VBD(1) NERQ: person-names(1), names (2)
Yago2s	rdfs:types(1), rdfs:subclass(1)	rdfs:types(1)
Yahoo! Categories	6 models	6 models

Table 7: Overview of the attributes utilized by both learners. Numbers in parenthesis denote the amount of models that integrated the corresponding property. Only groups containing elements incorporated into at least one models are shown.

pairs with higher MRR, we can say that the right class was closer to make it to the top. Recall that the reward obtained by a misclassified query is inversely proportional to the position where the correct label is. As a logical consequence, the black diagonal represents all instances obtaining an MRR score equal to one, that is to say when both the actual and assigned -highest ranked- labels coincide.

Both axes represent the twenty-six categories in congruence with their distribution in the corpus (see table 3). They are shown in decreasing order. This heat map highlights that the ranking gets worse when few training material is available. See, for instance, the white and light grey obtained by the smaller three categories: “*News & Event*”, “*Environment*” and “*Dining Out*”. Therefore, additional data is necessary to mitigate this effect, and hence to improve the performance for these categories. Here, we also envisage that semi-supervised approaches can aid by increasing the number of examples in these categories, or by benefiting from the content of previously clicked pages.

5.2. Feature Analysis

Figure 7 contrasts the features integrated into Winnow2 and MaxEnt models. Broadly speaking, three groups of characteristics shown to be the most discriminative for this task: a) the categories of previously clicked Yahoo! Answer pages; b) semantic analysis supplied by WordNet, collocations and ESA; and more important c) different formulations of the bag-of-words.

Yahoo! Answers categories. Despite of the substantially larger amount of available attributes, our outcomes corroborate our earlier findings that categories of previously clicked Yahoo! Answers pages are key to recognize the semantic fingerprint of question-like search queries. In particular, six Winnow2 and six MaxEnt models profitted from this attribute. Essentially, our experiments show that the larger the context the greater the contribution of this attribute. In

light of this result, we conjecture that enriching these models with the content embedded in previously visited pages might cooperate on enhancing the classification rate. In fact, we envisage the positive impact of all visited pages instead of only those from Yahoo! Answers.

WordNet and Collocations. MaxEnt models capitalized on several semantic relations, for instance selected collocations indicate that they are useful for disambiguating the semantics of the search query. A good example is given by finding the words "capital" and "resources" together within the same context. They are related nouns, and their co-occurrence is a good indicator for the presence of a Business & Finance question, whereas the semantics of each isolated word is ambiguous. Along the same lines, prepositions are useful for discriminating phrasal verbs, for instance the different semantics of the verb "get". By the same token, WordNet hypernyms and hyponyms were also useful. For instance, "market" is a kind of "activity", thus the presence of both words is a cue for Business & Finance questions. In summary, the presence of these relations specializes the semantics of its terms, hence assisting in recognizing their underlying topic.

Bag-of-Words. The outcomes, sketched in figure 7, point out to the fact that BoW characteristics are pivotal in this task. Interestingly enough, MaxEnt selected two variations regardless of the amount of context considered in the models: a) the alternative composed of lemmatized terms; and b) the one comprising raw terms without stop-words. Give this empirical result, we can draw two conclusions: a) terms with and without their morphological information are needed to semantically categorize queries properly; and b) one the one hand, the first BoW vector contains stop-words, while the other does not, meaning that this information becomes redundant. The effectiveness demonstrated by mixing lemmata and raw terms leads to think that generalizations are good at tackling data-sparseness and the lack of context provided by search queries, while raw terms (bearing morphological information) aid in capturing the specifics of each semantic category, and therefore in increasing performance.

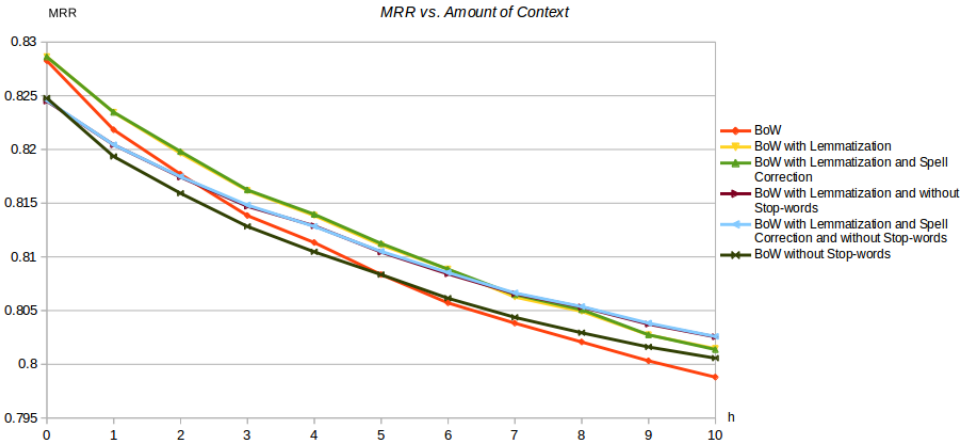

Figure 2: Comparison of MaxEnt models constructed on top of distinct bag-of-words (BoW) approaches. Results are shown for different levels of context ($h = 0 \ldots 10$).

In depth, figure 2 juxtaposes the outcomes reaped by each of the six models attempting distinct BoW variations. More specifically, this figure shows their performance accounting for context awareness at different levels, and for no extra features apart from the corresponding bag-of-words variation. There are some key aspects worth highlighting here. Our results indicate that a cost-efficient solution consists in building a traditional bag-of-words model from the target query only. They also signal that making allowances for the lemmata of its terms brings about a slight improvement. Our experiments show that benefiting from lemmata assists in finishing with the highest MRR score (0.8286). That is to say, regardless of the variation, the performance of the BoW model systematically drops as long as more preceding queries are taken into consideration when building the vectors. In a nutshell, this finding underlines the fact that in order to enhance performance, it is necessary not only to enlarge the context, but also to amalgamate different BoW approaches with additional features that can effectively capitalize on this enlargement. All in all, these figures support the valuable contribution of our proposed linguistically-motivated features.

Further, figure 2 shows that models considering stop-words outperform their counterparts without stop-words until some significant context is provided. Furthermore, our results also emphasizes the negligible effect of spell correction in enhancing the performance of BoW models. For instance, there is almost no difference between the lematized BoW with and without spell correction (green and yellow lines in figure 2). In conclusion, our results point out to the fact that a) stop-words and lemmatization cooperate on tackling the data-sparseness which characterizes short texts like search queries, however stop-words worsen the performance when larger contexts are available; and b) exploiting this linguistic processing proven to be much more effective than directly adding more context (raw terms) to the classification models.

Enriching the Lemmatized BoW with ESA. In substance, our six BoW models proven that this straightforward context mining approach is detrimental. Conversely, our experimental results reveal that facilitating Explicit Semantic Analysis for modelling the query context shows to be extremely helpful for increasing the classification rate. In detail, figure 2 highlights the results accomplished by MaxEnt models built via fusing two attributes: a) the lematized BoW of the target query; and b) the concept space provided by Explicit Semantic Analysis. This graph contrasts the outcomes obtained by taking into account different amount of context (h) and different concept vector lengths (p). On its base, contour lines for five levels of MRR were added.

Essentially, contour lines do not intersect and are scattered over separate regions of the space delimited by h and p, this means these levels of performance can be discriminated on the grounds of these parameters. Contrary to the BoW models, it unveils that best results (MRR ≈ 0.842) are obtained by mining contextual evidence from all ten historical queries, and to be more precise, a short vector composed of three to eight concepts is required in this case.

Next, the second level of performance (MRR ≈ 0.840) is achieved by means of harvesting at least six preceding queries. Here, the length of the concept vector decreases in consonance with increase in the amount of available prior queries, more specifically this vector can be shortened from three to one component depending on whether or not six or ten prior queries are available. Put differently, to some extent, the lack of contextual evidence can be

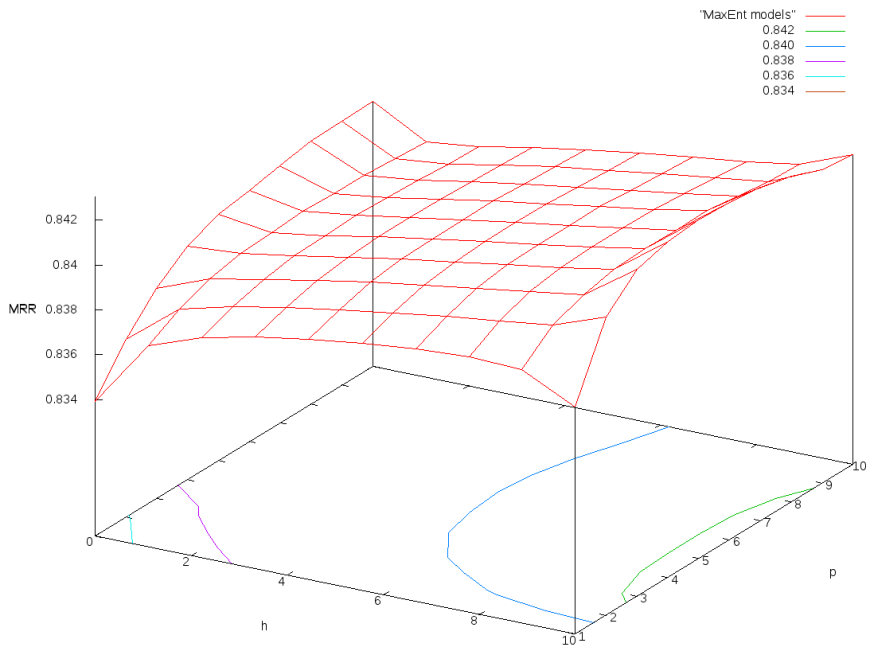

Figure 3: Impact of Explicit Semantic Analysis when combined with the Lemmatized bag-of-words (BoW) of the target query. Results are plotted for different values of h and p.

tackled by extending ESA vectors. Eventually, the proximity of the contour lines corresponding to the lowest three levels indicates that substantial gains can be made by exploiting the first three preceding queries. Specifically, these figures also suggest that when few context is available, it is a better alternative to incorporate only few, but top-ranked, concepts into ESA vectors. At large, the contribution of ESA becomes less dependent upon the length of the vector as long as the length of the transaction increases.

6. Conclusions and Future Work

This paper contributes to the research into community question-answering by enhancing the semantic classification of question-like search queries. This way the retrieval of related questions and answers from the community archives might be improved by matching manually entered question categories (by their members at posting time) with automatically determined classes for question-like queries (e.g., submitted in the search box of the platform or in a web search engine). More precisely, our work focuses its attention on effectively exploiting semantic cues yielded by preceding queries within the same question-like search session.

We discover meaningful discriminative properties by carrying out experiments on a large-scale dataset acquired automatically. By and large, our empirical outcomes indicate that the semantic processing provided by WordNet and collocation dictionaries is an important contributor to the betterment of the performance. But more relevant to mine the contextual evidence embedded in the session, it is exploiting the semantic concept space determined by Explicit Semantic Analysis. In effect, the contribution of this vector becomes less dependent upon its length as long as a larger

context is available. We also found out that lemmatization is pertinent to bag-of-words models, and that prior queries hurt their performance.

Our results also reveal that this task this is not hopelessly loss as extra sources of context might be exploited such as the content of previously visited pages. In reality, we deem that cost-efficient multilingual solutions can be implemented as ESA and BoW can be straightforwardly computed for several languages. Also, as a future work, we envision the use of connection across the search click graph and semi-supervised learning for tackling the data-sparseness caused by less frequent semantic categories. Lastly, we also envisage that ensemble methods can contribute to ameliorate the classification rate.

7. Acknowledgements

This work was partially supported by the project Fondecyt "*Bridging the Gap between Askers and Answers in Community Question Answering Services*" (11130094) funded by the Chilean Government.

GN: I probably will also add an acknowledgement here

[1] S. Zhao, H. Wang, C. Li, T. Liu, Y. Guan, Automatically generating questions from queries for community-based question answering, in: Proceedings of 5th International Joint Conference on Natural Language Processing, 2011, pp. 929–937.

[2] J. D. Denis Savenkov, Wei-Lwun Lu, E. Agichtein, Relation extraction from community generated question-answer pairs, in: NAACL 2015 Student Research Workshop, to appear, NAACL, 2005.

[3] F. M. Harper, D. Moy, J. A. Konstan, Facts or friends?: distinguishing informational and conversational questions in social q& a sites, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '09, ACM, New York, NY, USA, 2009, pp. 759–768.

[4] A. Rechavi, S. Rafaeli, Knowledge and social networks in yahoo! answers, 2013 46th Hawaii International Conference on System Sciences 0 (2012) 781–789.

[5] C. Barr, R. Jones, M. Regelson, The linguistic structure of english web-search queries, in: Conference on Empirical Methods in Natural Language Processing, EMNLP, 2008, pp. 1021–1030.

[6] W. K. Hanna, A. S. Aseem, M. Senousy, Issues and challenges of user intent discovery (uid) during web search, I.J. Information Technology and Computer Science 07 (2015) 66–76.

[7] A. Tamura, H. Takamura, M. Okumura, Classification of multiple-sentence questions, in: Natural Language Processing–IJCNLP 2005, Springer, 2005, pp. 426–437.

[8] A. Tamura, H. Takamura, M. Okumura, Classification of multiple-sentence questions, IPSJ 47 (6) (2006) 1954–1962.

[9] A. Figueroa, G. Neumann, Exploiting user search sessions for the semantic categorization of question-like informational search queries, in: Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013, 2013, pp. 902–906.
URL `http://aclweb.org/anthology/I/I13/I13-1115.pdf`

[10] D. E. Rose, D. Levinson, Understanding user goals in web search, in: WWW '04: Proceedings of the 13th international conference on World Wide Web, 2004, pp. 13–19.

[11] X. Yin, S. Shah, Building taxonomy of web search intents for name entity queries, in: World Wide Web Conference Series, 2010, pp. 1001–1010. doi:10.1145/1772690.1772792.

[12] X. Xue, X. Yin, Topic modeling for named entity queries (2011) 2009–2012doi:10.1145/2063576.2063877.

[13] J. C. K. Cheung, X. Li, Sequence clustering and labeling for unsupervised query intent discovery (2012) 383–392doi:10.1145/2124295.2124342.

[14] A. Figueroa, Exploring effective features for recognizing the user intent behind web queries, Computers in Industry 68 (2015) 162–169. doi:10.1016/j.compind.2015.01.005.
URL http://dx.doi.org/10.1016/j.compind.2015.01.005

[15] J. Guo, G. Xu, X. Cheng, H. Li, Named entity recognition in query, in: Research and Development in Information Retrieval, 2009, pp. 267–274. doi:10.1145/1571941.1571989.

[16] A. Eiselt, A. Figueroa, A two-step named entity recognizer for open-domain search queries, in: Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013, 2013, pp. 829–833.
URL http://aclweb.org/anthology/I/I13/I13-1101.pdf

[17] J. Du, Z. Zhang, J. Yan, Y. Cui, Z. Chen, Using search session context for named entity recognition in query, in: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR, 2010, pp. 765–772.

[18] D. Gayo-avello, A survey on session detection methods in query logs and a proposal for future evaluation, Information Sciences 179 (2009) 1822–1843. doi:10.1016/j.ins.2009.01.026.

[19] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, S. Vigna, The query-flow graph: model and applications, in: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008, 2008, pp. 609–618. doi:10.1145/1458082.1458163.

[20] A. Kotov, P. N. Bennett, R. W. White, S. T. Dumais, J. Teevan, Modeling and analysis of cross-session search tasks, in: Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011, 2011, pp. 5–14. doi:10.1145/2009916.2009922.
URL http://doi.acm.org/10.1145/2009916.2009922

[21] B. Hu, Y. Zhang, W. Chen, G. Wang, Q. Yang, Characterizing search intent diversity into click models, in: Proceedings of the 20th international conference on World wide web, ACM, 2011, pp. 17–26.

[22] Z. Liao, Y. Song, L.-w. He, Y. Huang, Evaluating the effectiveness of search task trails, in: Proceedings of the 21st international conference on World Wide Web, ACM, 2012, pp. 489–498.

[23] A. Kustarev, Y. Ustinovsky, P. Serduykov, Measuring usefulness of context for context-aware ranking, in: Proceedings of the 21st international conference companion on World Wide Web, ACM, 2012, pp. 551–552.

[24] B. Xiang, D. Jiang, J. Pei, X. Sun, E. Chen, H. Li, Context-aware ranking in web search, in: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, ACM, 2010, pp. 451–458.

[25] H. Wang, Y. Song, M.-W. Chang, X. He, R. W. White, W. Chu, Learning to extract cross-session search tasks, in: Proceedings of the 22nd international conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2013, pp. 1353–1364.

[26] H. Cao, D. H. Hu, D. Shen, D. Jiang, J. tao Sun, E. Chen, Q. Yang, Context-aware query classification, in: Research and Development in Information Retrieval, 2009, pp. 3–10. doi:10.1145/1571941.1571945.

[27] A. Broder, A Taxonomy of Web Search, in: SIGIR Forum 36:3-10, 2002.

[28] C.-Y. Lin, Automatic question generation from queries, in: Proceedings of Workshop on the Question Generation Shared Task and Evaluation Challenge, 2008, pp. 929–937.

[29] A. Figueroa, G. Neumann, Learning to Rank Effective Paraphrases from Query Logs for Community Question Answering, in: AAAI 2013, 2013.

[30] A. Shtok, G. Dror, Y. Maarek, I. Szpektor, Learning from the past: answering new questions with past answers, in: Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012, 2012, pp. 759–768. doi:10.1145/2187836.2187939.
URL http://doi.acm.org/10.1145/2187836.2187939

[31] Y. Liu, S. Li, Y. Cao, C.-Y. Lin, D. Han, Y. Yu, Understanding and summarizing answers in community-based question answering services, in: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Coling 2008 Organizing Committee, Manchester, UK, 2008, pp. 497–504.
URL http://www.aclweb.org/anthology/C08-1063

[32] A. Figueroa, G. Neumann, Category-specific models for ranking effective paraphrases in community question answering, Expert Syst. Appl. 41 (10) (2014) 4730–4742. doi:10.1016/j.eswa.2014.02.004.
URL http://dx.doi.org/10.1016/j.eswa.2014.02.004

[33] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, H. Li, Context-aware query suggestion by mining click-through and session data, in: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2008, pp. 875–883.

[34] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, the Journal of machine Learning research 3 (2003) 993–1022.

[35] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 1999, pp. 50–57.

[36] E. Gabrilovich, S. Markovitch, Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis, in: International Joint Conference on Artificial Intelligence, 2007, pp. 1606–1611.

[37] E. Gabrilovich, S. Markovitch, Wikipedia-based semantic interpretation for natural language processing, J. Artif. Int. Res. 34 (1) (2009) 443–498.

[38] S. Petrov, D. Das, R. T. McDonald, A universal part-of-speech tagset, CoRR abs/1104.2086.
URL http://arxiv.org/abs/1104.2086

[39] J. Morris, G. Hirst, Lexical cohesion computed by thesaural relations as an indicator of the structure of text, Comput. Linguist. 17 (1) (1991) 21–48.
URL http://dl.acm.org/citation.cfm?id=971738.971740

[40] M. A. Hearst, Texttiling: Segmenting text into multi-paragraph subtopic passages, Computational Linguistics (1997) 33–64.

[41] A. Eiselt, A. Figueroa, A two-step named entity recognizer for open-domain search queries, in: In IJCNLP, 2013, pp. 829–833.

[42] V. Lopez, C. Unger, P. Cimiano, E. Motta, Evaluating question answering over linked data, Web Semantics: Science, Services and Agents on the World Wide Web 21 (0).

[43] G. Andrew, J. Gao, Scalable training of l1-regularized log-linear models, in: Proceedings of the 24th International Conference on Machine Learning, ICML '07, ACM, New York, NY, USA, 2007, pp. 33–40. doi:10.1145/1273496.1273501.
URL http://doi.acm.org/10.1145/1273496.1273501

[44] Y. Tsuruoka, J. Tsujii, S. Ananiadou, Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1, Association for Computational Linguistics, 2009, pp. 477–485.

[45] N. Littlestone, Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm, in: Machine Learning, 1988, pp. 285–318.

[46] E. M. Voorhees, et al., The trec-8 question answering track report., in: TREC, Vol. 99, 1999, pp. 77–82.

[47] P. Devijver, J. Kittler, Pattern Recognition: A Statistical Approach, Prentice Hall, 1982.