Integrating Heterogeneous Sources for Predicting Question Temporal Anchors across Yahoo! Answers

Alejandro Figueroa[a,*], Carlos Gómez-Pantoja[a], Günter Neumann[b]

[a]*Departamento de Ciencias de la Ingeniería, Facultad de Ingeniería, Universidad Andres Bello, Antonio Varas 880, Santiago, Chile*
[b]*DFKI GmbH, Stuhlsatzenhausweg 3, Campus D3_2, D-66123 Saarbrücken, Germany*

Abstract

Modern Community Question Answering (CQA) web forums provide the possibility to browse their archives using question-like search queries as in Information Retrieval (IR) systems. Although these traditional IR methods have become very successful at fetching semantically related questions, they typically leave unconsidered their temporal relations. That is to say, a group of questions may be asked more often during specific recurring time lines despite being semantically unrelated. In fact, predicting temporal aspects would not only assist these platforms in widening the semantic diversity of their search results, but also in re-stating questions that need to refresh their answers and in producing more dynamic, especially temporally-anchored, displays.

In this paper, we devised a new set of time-frame specific categories for CQA questions, which is obtained by fusing two distinct earlier taxonomies (i.e., [29] and [50]). These new categories are then utilized in a large crowd-sourcing based human annotation effort. Accordingly, we present a systematical analysis of its results in terms of complexity and degree of difficulty as it relates to the different question topics[1].

Incidentally, through a large number of experiments, we investigate the effectiveness of a wider variety of linguistic features compared to what has been done in previous works. We additionally mix evidence/features distilled directly and indirectly from questions by capitalizing on their related web search results. We finally investigate the impact and effectiveness of multi-view learning to boost a large variety of multi-class supervised learners by optimizing a latent layer build on top of two views: one composed of features harvested from questions, and the other from CQA meta data and evidence extracted from web resources (i.e., snippets and Internet archives).

Keywords: Multi-view learning; Transfer learning; Question classification; Natural language processing; Intelligent information retrieval; Web mining;

1. Introduction

This paper studies temporal facets across user generated questions in Community Question Answering web services, like Yahoo! Answers[2], Stack Exchange[3] or Quora[4]. In such social web forums, people get the possibility to post questions of any kind with the expectation that other community members will provide good answers. If the asker is satisfied with some of these answers, he or she can provide feedback by explicitly marking the best answer. Since questions are now answered, these may be closed and archived so that they are available in the future, e.g., as potential answer sources for new (same or similar) question posts. On the other hand, the asker feedback also has benefits for the answerer, because the more "best answers" he or she produces the more reputation this person may amass in

[1]The new annotated corpus will be made publicly available upon acceptance under http://something.here.com.

[*]Corresponding author; phone: +56 (2) 27703795

Email addresses: alejandro.figueroa@unab.cl (Alejandro Figueroa), carlos.gomez.pantoja@unab.cl (Carlos Gómez-Pantoja), neumann@dfki.de (Günter Neumann)

[2]https://answers.yahoo.com/

[3]https://stackexchange.com/

[4]https://www.quora.com/

the CQA forum. In this traditional scheme, archived questions are re-used based on their semantic connections with newly published questions. That is to say, this search for related content is aimed predominantly at finding "more like this" at the expense of its diversity (i.e., semantically dissimilar or loosely semantically related questions). Needless to say, presenting diverse outputs helps to stir up the interest of community members to acquire knowledge by exploring new topics. To illustrate how temporal anchors can assist in bridging the diversity gap, consider the following pair of Christmas-anchored semantically-dissimilar questions "*After leaving Bethlehem, to which country did Joseph, Mary, and Jesus travel?*" and "*How to cook Christmas turkey?*". In reality, temporal anchors do not cooperate solely on fetching strongly related content (e.g., more Christmas cooking-recipes), but also and more importantly, they play a pivotal role in discovering interesting, which otherwise would be unrelated, material.

In effect, it is vital for boosting the diversity and dynamicity of these platforms to exploit their semantical richness, especially taking into account that their increasing popularity stems from allowing their users to get fast and accurate answers to complex natural language questions, directly from a community [7, 51]. To exemplify their semantic variety, Yahoo! Answers distinguishes between 26 top–level categories (see also Table 3, page 7). So far, Yahoo! Answers allows to filter their search results by categories or by time, where time here means the moment when questions were archived. However, besides these sorts of extensions, the exploration of CQA repositories is still mainly text-based and surface oriented.

Another way in which the identification of temporal anchors can help sites and search engines (that return CQA answers as part of their search results) to manage their repositories is filtering out –or devising strategies to deal with– outdated content. E.g., questions asked during repeated sport events like the Olympic Games or World Soccer Championships (e.g., "*Who will win Chelsea or Arsenal?*"). It can also assist in coping with questions which usually receive a high impact for a short period of time like those happening during a natural disaster or the marriage of famous people (e.g., "*Who killed Anthony Scalia?*"). Broadly speaking, the benefit of adding temporal categories to the archived meta data may lead to better member experience.

Currently, there are two viewpoints for temporality across CQA sites: a) a measure of the usefulness of the answers[50]; and b) the recurrent attention given to questions during different time-frames[29]. The purpose of this work is to fuse these two approaches in order to achieve a broader perspective of the concept of question temporality and to carry out substantial experiments on basis of a rich and diverse feature set. In particular, we systematically take into account the large set of topic categories provided by Yahoo! Answers in order to investigate how different is the complexity of the identification of these temporal anchors across distinct topics, and if so, whether this behaviour is the same for humans and for machines. For this purpose we develop a much larger human annotated corpus than introduced in previous work, and use it in a crowd-sourcing system with up to fourteen workers. The new corpus is based on Yahoo! Answers (text of questions and their answers, profile information and meta data) and does not depend on additional sources like search engine web clicks[5]. In summary, our main contributions are:

- We propose a new set of time-frame specific categories, which are obtained by fusing the different categories from [50] and [29].
- We describe the process and the results of a large crowdsourcing based human annotation effort of a new question data set. We systematically analyse the complexity and degree of difficulty of human annotation of questions coming from different topics, and what we can learn by this analysis about the difficulty of the corpus labelling process.
- We create a high quality new corpus of Yahoo! Answers questions and answers containing 6683 questions labeled manually with the new set of time-frame specific categories.
- Through a large number of experiments, we investigate the effectiveness of a wide variety of linguistic features compared to what was done in previous work.
- Moreover, we are also mixing evidence/features distilled from heterogeneous resources viz. directly and indirectly from questions implying web searches and Internet archives.

[5] Our annotated corpus will be publicly available upon acceptance under http://something.here.com

- Based on these two views, we investigate the impact and effectiveness of multi-view learning to boost a large variety of multi-class supervised learners.

The major outcomes of our research can be summarized as follows. Firstly, using Sequential Forward Floating Search (SFFS) [52] as baseline for multi-view learning, we observed that linguistic information is substantial for identification of temporal anchors, and that web search is substantial for identifying relevant text fragments (see sec. 4.1). We found out that humans and machines show different degree of difficulties when labeling questions from diverse topics. A topic that is easy to label by a human, might be difficult to label by a machine, and vice versa. Thus, at least in this task, the interpretability of machine decisions might be hard to achieve. Secondly, using a Dual version of SFFS improves the classification performance, but on different feature combinations compared to SFFS (see section 4.2). For example, information from profiles and meta data seems to be more valuable for Dual SFFS than for SFFS. However, we also observed that the degree of difficulty in the assignment of labels to questions is similar to the observations we made for SFFS. Furthermore, independently of the chosen multi-view learner, same topics seem to have same difficulty degrees. Thirdly, introducing and exploring Category-based Transfer Learning (CbTL) ensembles in the context of CQA as an alternative to Dual SFFS were less successful as expected (see sec. 4.3). Actually, our intuition that distinct classifiers should be utilized for different target inputs could not be verified by the results of our experiments, since they were even lower than the results of SFFS.

The article is structured as follows. We first present a brief overview of related work in section 2, before we present the technical background of our work in section 3. This covers details about the acquisition and the annotation process of the corpus in subsections 3.1 and 3.2, a characterization of the explored features 3.3, and a description of the multi-class supervised learners and their embedding into multi-view and transfer learning strategies, cf. subsection 3.4. In section 4 the experiments are introduced, followed by a detailed description and analysis of the results obtained for the baseline (cf. subsection 4.1), Dual SFFS (cf. subsection 4.2) and transfer learning (cf. subsection 4.3). Finally, section 5 summaries the main aspects of the paper and outlines some future directions.

2. Related Work

Community Question Answering (CQA). One recent research trend focuses on the recognition of question similarities, e.g., as a means of detecting and ranking similar questions, e.g., [28, 53, 56]. Also, research into CQA sites is paying attention to the recognition of question paraphrases and question answer ranking/retrieval [51], to the detection of communities as well [41, 44]. In [5] different measures used to evaluate question quality in CQA websites are surveyed. They focus on question related features and showed that question features most frequently used in research into predicting question quality were tags and terms, length of the question, the presence of an example and user reputation. In [60] a large review of CQA web forums is described, where they point out in the discussion section that user spatial, temporal, and social context in CQA should play a more significant role especially in mobile devices. Indeed, only very few work has been published about the aspect of temporality in CQA forums, cf. [29] for temporality in questions, and [50] and [69] for temporality amongst answers. Still a main open research question is about the identification and definition of appropriate time-frame taxonomies, and the question of how to obtain high-quality data annotations. This is exactly one aspect and motivation of the proposed approach described in this paper.

In details, [50] were the first who introduced the concept of temporality as a measure of the usefulness of the answers provided on the questions asked in CQA web forums. They focused on that part of temporality, where the answer to a question is quite likely to expire or become obsolete. This might happen for questions where the point of time is only referenced implicitly. For example, for the question "What day is Thanksgiving?" the best answer found in the archive is "22nd November", which is correct for the year 2007, but not automatically for later years. Thus, a user-friendly CQA should not consider this answer for the same question posted in the year 2011. As a result, [50] defined a set of five different time-frame specific categories (permanent, long-/medium-/short-duration, other) and sampled and manually annotated a small data set of 100 questions from Yahoo! Answers with these categories to learn a classifier.

A recent extension of this line of research is described in [29]. They focused on the recurrent attention given to questions during different time-frames. In particular they utilized the relationship between search logs and Yahoo! Answers pages connected via Web user clicks as a source for the analysis of temporal regularities of user interests across CQA questions. In contrast to [50], they focus on when likely a question will be asked (or asked again) rather

than when the answer of a question will be outdated. As a result they defined four time-frame specific categories (permanent, periodic, trend, and others) and automatically created a large development data set of 35.000 questions. These questions are verified manually (on basis of binary decisions), and then later used to evaluate the performance of different supervised classifiers.

In the context of factoid QA systems[6], [24] have recently presented a neural architecture that encodes not only the content of questions and answers, but also the temporal cues in a sequence of ordered sentences which gradually remark the answer. Some earlier work have focused on the identification and splitting of complex temporal questions for question answering systems, e.g., [34], [48] and [55]. However, they focused on the identification and analysis of date expressions in questions like "Who won the Nobel Prize in physics before 1970?", where our work focuses on the classification of questions with respect to certain time-frames, i.e., when will a question more likely be raised. A classification of Question Answering Systems (QASs) based on explicitly identified criteria like application domains, questions, data sources, matching functions, and answers is presented in [46]. They present a systematic survey of major QAS and their results suggest that temporal aspects have not yet been in the forefront of QAS research. In a similar fashion, [33] discuss in their QAS survey only simple When-questions which starts with the keyword "When" under the aspect of temporality.

Web Search and Temporality. [29] utilize the relationship between Web search logs and Yahoo! Answers pages connected via user clicks as a source for the analysis of temporal regularities of user interests across CQA questions. They define three main types of temporally anchored questions: spiky or bursty, periodic and permanent. According to [61], a query burst is a, frequently short, period of heightened interest of users on a particular topic, which brings about higher frequencies of related search queries. Contrary to spiky queries, this period of heightened interest is recurrent and very predictable in the event of periodic requests, while permanent queries are often likely to have very small variations in their frequencies. They also characterize stable queries by very small variations over time in a metric called burst intensity.

In a survey paper of temporal web search experience, results of [36] suggest that an interplay of seasonal interests, technicality of information needs, target time of information, re-finding behaviour, and freshness of information can be important factors for the application of temporal search. Our findings summarized in this paper somewhat extend these results to the domain of CQA. An interesting approach that maps the contents of a document to a specific time period is introduced in [57]. The idea is to treat documents and years as nodes which are connected by intermediate Wikipedia concepts related to them. Identifying this time period associated with the document can be useful for various downstream applications such as document reasoning, temporal information retrieval, etc. More generally, `https://en.wikipedia.org/wiki/Temporal_information_retrieval` gives a good overview of relevant other aspects explored in the field of temporal information retrieval.

Time Expression Recognition. It is a fine-grained task aimed at automatically identify time expressions from texts, and normally, it does not only encompass the recognition, but also the normalization of these expressions. Take for instance, [73] discovered that time expressions are formed by loose structures, and their words differentiate them from common text. In general, most strategies for time expression recognition can be categorized into rule-based [13, 74] and learning-based methods [3, 6, 30, 39].

Multi-view machine learning. Multi-view machine learning is a rapidly growing direction in machine learning with well theoretical underpinnings and great practical success [62]. It is concerned with the problem of machine learning from data represented by multiple distinct feature sets. Different strategies have been proposed ranging from unsupervised to supervised methods. They are further classified into three groups based on the distinct views (e.g., redundant or collaborative) they have on a given feature set: co-training, multiple kernel learning, and subspace learning [70]. Our approach falls into the last group as it constructs a latent subspace on top of two distinct collaborative views cf. also section 3.4. More precisely, we present a multi-view strategy based on ensemble learning, and one based on transfer learning. The goal of ensemble learning is to use multiple models (e.g., classifiers or regressors) to obtain a better predictive performance than could be obtained from any of the constituent models [71]. The goal of transfer

[6] In such a QA system a question usually requests a single fact as answer, e.g., "Elon Musk" is the answer to the question "Who is the CEO of Tesla?". Note that this is in contrast to the question and answer style in CQA which are in general non-factoid questions.

learning is to transfer knowledge learned in one or more source tasks to a related target task to improve learning [14]. A recent survey of ensemble learning strategies in the context of expert finding for CQA is presented in [72]. The benefit of transfer learning for fact-oriented question answering (QA) of models trained on a different large, fine-grained QA dataset is demonstrated in [45].

Crowd-based data annotation. Crowdsourcing is considered as a cheap, fast and reliable mechanism for gathering labels. [58] discuss the use and benefit of crowdsourcing in the context of Natural Language Processing. They argue that, in general, volunteer-supplied data or data supplied through Amazon Mechanical Turk (AMT) is more plentiful but noisier than expert data. Consequently, [1] consider the question of how many workers are needed to obtain high quality labels. Our approach follows the ideas presented in that paper and we are describing the outcomes of experiments in the context of CQA using up to fourteen workers, see also subsection 3.2. For a general survey of quality control in crowdsourcing see [20].

3. Integrating Heterogeneous Sources for Predicting Question Temporal Anchors across Community Question Answering Platforms

3.1. Corpus Acquisition

The first step consists in acquiring a working corpus for our study. For this purpose, we designed a crawler to navigate through the Yahoo! Answers site from September 2015 to January 2016. According to the dynamic of this service, each time a new question is posted, community members are obliged to categorize it in accordance with their three-level taxonomy. In this system, top-level classes are broad and embrace a constantly growing massive amount of questions and answers. On the flip side, most fine-grained classes at the bottom (third-level) are more specific, therefore they have narrow coverage and seldom get new questions.

With this in mind, our crawler was devised to navigate through questions posted across categories embodied only at first two levels. When browsing each category page, it retrieves the top ten questions displayed by the platform. Note also that each of these category pages was visited several times during this time frame in order to increase the volume of its questions, since new questions were surely posted during these five months of crawling, and these might appear within the top ten hits. As a logical consequence, this revisiting policy assists in accumulating sets of instances that encompass a wide variety of topics. In total, we gathered almost 370,000 question pages and all their titles, bodies and answers were stored accordingly.

However, this crawler was not designed to filter downloaded Yahoo! Answers pages by their language. Thus we capitalized on a language detector[7] for singling out all questions and answers written predominantly in English. After filtering, we retained ca. 180,000 questions in English. Subsequently, we randomly selected 265 questions from each of the 26 top-level categories, and manually removed spurious instances afterwards. All in all, we ended up with 6683 questions as our study collection.

3.2. Corpus Annotation

One of the contribution of this work is fusing two taxonomies proposed in two distinct earlier studies, i.e., [29] and [50]. In the first place, we consider the viewpoint of temporal anchors developed by [29], defined as the period of attention a question might grab. Second, influenced by the study of [50], our proposal also takes into account the timeframe where its answers are valid, when outlining this taxonomy. In detail, our proposed merge is shown in Table 1. In order to manually assign these temporal anchors to each question in our study corpus, we followed the approach of [1]. A key feature of this method is that it models the annotation process as a stylized crowd-sourcing system that operates in rounds[8]. In each of these rounds, the system isolates one question and asks an assessor to submit his/her judgment and then gets paid for the work. Since this crowd-sourcing system needs to produce a final answer for each question, it can adaptively decide for each element the amount of annotators to ask for judgments.

Basically, this algorithm requires a stopping rule to decide whether or not to stop asking for judgments given a question. After stopping, it additionally requires a selection rule that allows to determine the final label from the

[7] code.google.com/archive/p/language-detection/

[8] Our annotated corpus will be publicly available upon acceptance under http://something.here.com

Anchor	Question	Answer
Periodic	The interest of the question conspicuously increases during determined and specific time frames.	Answers can be reusable. In other words, same answers can be used when a new occurrence of the event/topic happens.
	Examples: *"How do you cook a Christmas Turkey?"*, *"What are good ideas for Valentines Day?"*, *"When is Yom Kippur?"*	
Spiky/Bursty	The interest for the question starts and dies abruptly. It captures great attention suddenly for a short period of time, and then this interest dies quickly.	Answers to these questions grab the attention for the short period of time that the question lives. Then, it is unlikely that they will be consulted later. Though answers might still be valid.
	Examples: *"When will Hurricane Sandy hit NYC?"*, *"Did Obama killed Scalia?"*, *"Who killed Osama Bin Laden?"* *"Will Trump win tonights SC primary?"*	
Permanent Recyclable/ Non-Recyclable	They can be fetched at any moment. The level of interest is on average constant and normally very low during any period of time. Mostly factoid questions.	Answers to these questions might or might not be reusable later. Questions might have multiple good answers. The core of the answers is factual info. They might be not reusable because the answer will expire or expired.
	Examples: *"How to make green beer?"*, *"How do you remove acne?"*, *"What is the capital city of the United States?"*, *"What is the time difference between Santiago and London?"*	
Multiple Spiky/Bursty	They behave like bursty questions, but repeatedly. However, the period between consecutive instances is undetermined.	Answers are not reusable. That is to say, answers to the previous occurrence are not useful for the new happening.
	Examples: *"Are you pro-life or pro-abortion?"*, *"Will the GOP win this election?"*, *"Are you for or against of gun control?"*, *"Who will win tonight Real Madrid or Barcelona?"*, *"How much did the stock market crashed yesterday?"*, *"How many red cards has Luis Suárez received this year?"*, *"Did Angelina Jolie and Brat Pitt get divorced?"*	
Trend/Drift	The interest for the question increases slowly, normally it reaches a plateau and then decreases slowly.	Answers are reusable, reaching a peak of attention. Later, the interest decays and it will be seldom retrieved.
	Examples: *"How do I install Windows 8?"*, *"How do I make furry nails?"*, *"How do you get an iphone 5s or 6 for CHEAP?"*	
Other	All instances that annotators deemed unfitted to all other categories.	

Table 1: Definitions of classes in the taxonomy of temporal anchors for questions proposed by our work.

collected judgments. A key advantage of this method is that it amalgamates both criteria in such a way that it reduces both the error rate and the annotation costs.

The underlying idea behind this adaptive mechanism is that some questions are very easy to label, therefore there is no need to ask for judgments to a large number of assessors, since most of these inputs will be redundant and will unnecessarily increase the overall tagging cost. Conversely, the labels of other elements are very difficult to determine, and for this reason, more judgment will be required to mitigate their annotation error rate. Put differently, less judges are needed to deal with easy questions, whereas more assessors with difficult questions. Here, the notion of easy/difficult is given by a reflection of the agreement of the majority, rather than of the sentiments of the assessors. More precisely, a question is hard to label if the distribution of its labels, provided by a group of assessors, is closer to even, whereas it is easy if an early strong bias towards an option is clearly observed.

In our annotation process, we assumed that all assessors are anonymous, i.e., we had no prior information on which judges are better than others, ergo all inputs have the same weight. Specifically, we accounted for diverse group of up to fourteen assessors per question including undergraduate students, mechanical turkers and professionals. According to [1], the stopping rule when more than two labels are available is given by:

$$Stop\ if\ V_{A^*(t),t} - V_{B^*(t),t} \geq C\sqrt{t} - \epsilon t \tag{1}$$

In this rule, t is the number of labels available for a question (i.e., $t = 2 \ldots 14$). $A^*(t)$ and $B^*(t)$ are the labels with the largest and second-largest amount of votes $V_.$, respectively. The selection rule chooses the most voted option as the final label, but if the stopping rule cannot be satisfied after the fourteenth judge, it randomly chooses according to the probability given by the vote distribution. In our annotation process, we experimentally set the parameters C and ϵ to 1.5 and 0.25, respectively.

This annotation method does not only balance the error rate with its inherent cost, but its outcome also aids in drawing interesting conclusions regarding the corpus prior to the experimental phase. Particularly, in 35.23% of our questions, the inputs of only the first two judges were required, since they agreed (see some samples of annotation in Table 2). The labels of four assessors were required solely for 8.64% of the elements within our collection. This means that one third of the instances required few (two) judges to be determined. In this group, we find 64% of instances fell

Category-Label/No. judges/Date	Question Tile and Body
Environment Spiky/Bursty/2/2016-01-20	**To global warming deniers,does this article prove global warming is true?** www.sciencedaily.com/releases/2016/01/160120115506.htm
Yahoo! Products Multiple Bursty/2/2013-02-22	**What happened to my yahoo page style?** Yahoo page style has changed can I get back to where it was before it changed?
Computers & Internet Drift/10/2015-09-23	**Can i just install windows 7 over Xp?** Is the any requirements?
Travel Periodic/2/2012-12-08	**What is Sevilla like in the spring?** Festivals, weather, anything else that is important too.
Yahoo! Products Other/4/2014-08-07	**POLL: It's been about 4 years since I was on here. Are any of my friends still on here?**
Dining Out Permanent Recyclable/ 4/2013-03-01	**Where can i find choclate covered strawberries in hyderabad?** Im craving for them like crazy... Can any one tell me where can i get choclate covered strawberries in hyderabad.... Im ready to go to any corner of hyderabad to find them... Please tell me where can i find them..
Travel Permanent Non-Recyclable/ 8/2015-11-01	**Which is better to Live west Hollywood or north Hollywood?** So in 3 years I am moving to California, I wanna go out there for school and to try and start modeling and im just trying to gather as much info as I can about north and west Hollywood(the school I wanna go to is in the heart of Hollywood)

Table 2: Samples of manually annotated questions.

Question Category	Average	%	Question Category	Average	%	Question Category	Average	%
Science & Mathematics	4.15 (0.24)	16.08	Sports	5.34 (0.32)	34.11	News & Events	5.97 (0.33)	26.27
Computers & Internet	4.39 (0.25)	21.88	Education & Reference	5.42 (0.27)	17.12	Games & Recreation	6.03 (0.31)	33.07
Cars & Transportation	4.84 (0.28)	22.27	Environment	5.55 (0.32)	28.02	Beauty & Style	6.32 (0.30)	21.18
Home & Garden	4.86 (0.25)	16.08	Arts & Humanities	5.63 (0.27)	20.78	Society & Culture	6.51 (0.31)	27.45
Consumer Electronics	4.88 (0.32)	35.94	Food & Drink	5.63 (0.27)	15.95	Pregnancy & Parenting	6.52 (0.26)	19.14
Local Businesses	4.92 (0.26)	18.87	Health	5.65 (0.28)	16.80	Social Science	6.62 (0.31)	29.02
Yahoo! Products	5.19 (0.28)	14.94	Dining Out	5.66 (0.31)	26.89	Entertainment & Music	6.86 (0.31)	25.49
Travel	5.21 (0.29)	25.58	Politics & Government	5.77 (0.32)	28.52	Family & Relationships	7.23 (0.24)	19.46
Business & Finance	5.31 (0.28)	22.48	Pets	5.88 (0.27)	16.67			

Table 3: Top-level question categories vs. the average number of judges needed to tag their questions. In parentheses, we find the respective standard deviation. The other % signals the fraction of elements requiring a final random decision.

into the time-frame category *Permanent Recyclable*. On the flip side, 25.31% questions required all fourteen assessors to submit their judgments. In 23.08% of the cases, the label still remained undetermined after the fourteenth judge due normally to two pretty tied options. In these cases, the selection was randomly drawn, accordingly.

From another angle, Table 3 shows the difficulty in the annotation process with respect to the question category in terms of both the average number of required assessors and the portion of labels randomly defined. The Pearson Correlation Coefficient (PCC) between both the average amount of judges and the portion set by random labels is 0.16, indicating a weak correlation. Overall, our analysis indicate that it is easier and cheaper to manually determine the temporal anchor of questions coming from categories such as *Science & Mathematics*, *Home & Garden* and *Yahoo! Products*. In juxtaposition, it is harder to manually assess the temporal anchor of elements derived from *Social Science*, *Entertainment & Music* and *Family & Relationships*. Roughly speaking, the average number of judges required by *Family & Relationships* doubles *Science & Mathematics*.

From another standpoint, Bursty/Spiky questions are prominently found across categories including *News & Events* (25.38%) and *Politics & Government* (16.84%); Multiple Bursty/Spiky within *Sports* (33.33%) and *News & Events* (19.05%); Trend/Drift in *Computers & Internet* (18.62%) and *Consumer Electronics* (18.09%); Periodic within Travel (12.35%) and Sports (11.11%). The remaining three temporal anchors are more evenly distributed across question categories, being Permanent Recyclable less frequent in *News & Events* (1.18%), while Permanent Non-Recyclable within *Politics & Government* (1.73%) and *Computers & Internet* (2.05%).

In addition, we ask assessors to provide general insights into why they decided to label some questions as Other as a means of gaining extra understanding on question temporality. Some of the interesting insights include:

- Assessors felt that some questions did not fit any class, though they could not provide any reason why they had

Question Category	Other (%)	Not Temporal Anchored (%)	Temporal Anchored (%)	Entropy (3)	Entropy (7)
Arts and Humanities	49.41	39.61	10.98	1.38	1.55
Business and Finance	37.6	47.67	14.73	1.45	1.62
Consumer Electronics	23.44	48.83	27.73	1.51	1.96
Education and Reference	39.3	49.03	11.67	1.4	1.53
Entertainment and Music	52.94	21.57	25.49	1.47	1.88
Health	34.77	59.77	5.47	1.2	1.25
Games and Recreation	43.97	34.63	21.4	1.53	1.89
Science and Mathematics	20.78	72.16	7.06	1.08	1.15
Beauty and Style	52.16	37.65	10.2	1.36	1.5
Sports	37.6	30.23	32.17	1.58	2.24
Social Science	49.02	38.82	12.16	1.4	1.6
Cars and Transportation	25	62.5	12.5	1.3	1.35
Dining Out	37.88	40.15	21.97	1.54	1.89
Food and Drink	32.68	58.75	8.56	1.28	1.43
Home and Garden	29.02	62.35	8.63	1.25	1.34
Local Businesses	34.34	48.3	17.36	1.48	1.63
Family and Relationships	69.26	20.62	10.12	1.17	1.33
News and Events	28.63	13.73	57.65	1.37	2.19
Pets	39.92	52.71	7.36	1.29	1.39
Politics and Government	27.73	34.38	37.89	1.57	2.12
Environment	25.29	44.36	30.35	1.54	2.06
Society and Culture	47.84	36.47	15.69	1.46	1.73
Travel	28.29	50	21.71	1.49	1.85
Computers and Internet	19.92	53.91	26.17	1.45	1.81
Pregnancy and Parenting	55.47	35.55	8.98	1.31	1.45
Yahoo! Products	26.05	60.15	13.79	1.34	1.6

Table 4: Label distribution across each question category. Into "Temporal Anchored" are clustered all five classes that identify some sort of time-dependency (e.g., Periodic, Spiky, Permanent Non-Recycle and Multiple Spiky). Conversely, under "Not Temporal Anchored", we find all instances tagged as Permanent Recyclable. Entropy(3) denotes the entropy by grouping our seven labels into the two broader groups plus Other, while Entropy(7) is calculated wrt. the original label set.

this feeling. On the flip side, they noted that some questions seemed to fit multiple categories.

- In the same spirit, judges pointed out questions that are intrinsically the same, but a slight change made them to have a markedly different temporal anchor. To illustrate, consider the pair "*How Whitney Houston died?*" (likely Bursty) and "*How JFK died?*" (probably Permanent Recyclable).

- Some questions were unintelligible, e.g., underspecified, linked to broken sites or their language was incorrectly guessed. Other questions were perceived as spurious (e.g., song lyrics). Some questions were deemed as unnecessary by the annotators, take for instance: "*Happy new year 2016 to everybody*".

- Lastly, judges felt that some questions and their answers were not reusable, in particular elements where their narrative targeted personal issues. They conceived these personal questions as a-temporal (e.g., asking about personal appearance).

Last but important, Table 4 compares the distribution of labels across different question categories. Here, Entropy (3) signals the entropy of the class distribution when putting questions together into three broader groups: Other, temporally and non-temporally anchored elements. Note that, in this case, the higher achievable entropy value is 1.585, and these broader groups provide insight into the impact of the temporally-anchored material on the distinct question categories. Also, it is worth highlighting that twelve out of 26 categories are very close to this maximum value (at least 90%). All things considered, temporal anchors are seldom found across *Science & Mathematics* and its content is highly-recyclable, while *Sports* and *Politics & Government* are the most evenly distributed. A very similar picture is found when computing the entropy wrt. the seven original classes (maximum value of 2.8). However, different temporal anchors are likely to be concentrated on different categories, for instance, Spiky is more easily found in *Politics & Government* where as Periodic in *Travel*.

How do i uninstall windows 10? (posted on 14th Oct. 2015)

Rank	Timeframe	Times saved
1	July 1, 2015 and November 29, 2017	257
2	September 12, 2016 and May 7, 2017	17
3	July 18, 2016 and November 22, 2017	15
4	July 30, 2016 and August 20, 2017	17
5	August 9, 2015 and May 21, 2017	38
6	August 11, 2016 and December 7, 2017	5
7	August 8, 2015 and July 16, 2017	114
8	January 14, 2016 and June 7, 2017	119
9	August 24, 2016 and December 21, 2016	16
10	July 31, 2015 and July 9, 2017	116

Table 5: Aggregated crawling dates harvested from the Internet Archive for the CQA question "*How do i uninstall windows 10?*". Entries are listed in agreement with the ranking given by StartPage. "Times saved" denotes the amount of crawls registered during the respective Timeframe.

3.3. Features

Broadly speaking, we constructed high-dimensional feature spaces by means of fusing two different sources of attributes: the web and community platform content.

With regard to the web, we profit from the StartPage[9] search engine for finding documents pertaining to each question on the web. For this purpose, we requested this engine ten hits for each question title. Since the language used in Yahoo! Answers is informal, and thus its content is sometimes riddled with typos, question titles were orthographically corrected by means of Jazzy[10] before submission. From each retrieved web snippet, we extracted its respective title, description and url, which were utilized for further processing. To be more exact, we capitalized on these extracted urls for retrieving the crawling dates registered by the Internet Archives (a.k.a. Way Back Machine[11]). Although, crawling dates are not indicative of interest, these timestamps can be used as a way of roughly estimating the starting point of a topic (if any). It is worth noting here that sometimes these timestamps match the respective period of interest. In addition, these can be used as a reference for detecting when the interest for a topic died, and therefore its pages ceased to exist. Take the example provided in Table 5, Windows 10 was officially released on July 29, 2015, and for this reason we can find that the earliest crawled pages date back to July 2015. Since there is some evidence that these web pages still exist, we can conjecture that this topic might still be of some interest.

As for features, we extracted from this view the number of registered crawls for each hit returned by StartPage. We additionally capitalized on the number of crawling dates that matches the day, the month and the year of the question. We also benefited from the web snippets for counting the number of times the question's day, month and year appear within their urls. The hosts of these urls were also perceived as features. Furthermore, we extract linguistic features from these web snippets by profiting from CoreNLP[12][43]. The following linguistic characteristics were computed independently from both its title and body:

- **Bag-of-words (BoW):** It was constructed by taking into account traditional raw term frequencies. We also built an alternative version via lemmatized terms.

- **Named-Entities (NER):** CoreNLP NER annotator recognizes named entities (i.e., person, location, organization and misc), numerical (i.e., money, number, ordinal and percent), and time entities (i.e., date, time, duration and set). For each of these entity classes, we constructed a BoW-like vector modelling the occurrence of each entity found across the snippet. Additionally, we counted the number of times the day, month and year of the question appears within the snippet. We also accounted for matches in the day of the week (i.e., Monday and Saturday) and year (i.e., 1-365), and also for the week (i.e., 1-52) in the year. Since this sort of temporal information rarely appear across snippet titles, merged counts were considered for this effect.

[9] www.startpage.com
[10] jazzy.sourceforge.net
[11] archive.org/web/
[12] stanfordnlp.github.io/CoreNLP/

All these counts were accumulatively computed from the first to the k snippet ($k = 1 \ldots 10$), in this way we intent to discover the best level of retrieval ranking necessary to make the best out of each property. It is worth emphasizing here that we normalized all date expressions in order to perform their corresponding matches (e.g., Aug, August and 08 were all mapped to 08). We also added as attributes the question day, month, year, hour, minute, am/pm, day of the week and year, the week in the year as well. Furthermore, we extracted several community meta-data characteristics, especially from the member card: gender, level, joining year, their points in the logarithmic scale, percentage of best answers, the number of answers and questions in the logarithmic scale, url hosts, and the number of sentences used in their self-descriptions. Furthermore, from these self-descriptions and the questions, we computed the following linguistic attributes:

- **Bag-of-words (BoW):** We split this traditional vector representation into distinct elements. First, we considered a BoW comprising only stop-words. We also made allowances for a BoW encompassing all true case modifications proposed by CoreNLP. We additionally took advantage of sentiment analysis for constructing a BoW for each sentiment level (i.e., using a five point Likert scale). We also constructed a BoW of lemmata for all terms that did not appear in their root form. We additionally built a BoW for each universal POS tag. We also constructed a BoW for all resolved pronoun references.

- **Named-Entities (NER):** We took into account a BoW for each named entity class. We additionally perceived as features the highest frequent entity and its respective class.

- **Parse Tree (PT):** We conceived as features the type of the first constituent and the frequency of each constituent class. Since it is conjectured that temporal phrases are compositional in nature [4], we expect to capture the temporal essence of questions that are more frequently manifested across certain kinds of constituents (e.g., ADJP). To exemplify this compositional nature, [38] claimed that temporal adjectives (e.g., new and later) are recurrent across subordinate clauses brought in by temporal markers including before and after.

- **Lexicalised Dependency Tree (DP):** Here, we profited from two BoWs. One composed of the root nodes, and the other one of the frequency of each relationship type. We also interpreted as features the level of the shallowest, average and deepest tree. The number of nodes at the first five levels of the tree. The minimum and maximum number of children of a node, and their respective averages. Simply put, some dependency types (i.e., tmod) aim at modifying the meaning of VPs or ADJPs by specifying a time.

- **HPSG parser[13]:** Overall, we used this parser for carrying out a deeper linguistic analysis on verbs [47]. We count passive/active verbs and auxiliaries (e.g., copular, have and modal), besides the amount of items falling into each potential tense (e.g., present, past and untensed) and different aspects (e.g., perfect and progressive). And across all sorts of terms, we counted kinds (e.g., noun/verb modifiers) and lexical entries (e.g.,[< ADVP >]ADJ-adv_superative_rule). In all six cases, we accounted additionally for the highest frequent item as attribute (e.g., voice, tense and type). We hypothesize that this sort of verb enrichment (e.g., tense and voice) will cooperate on recognizing some kinds of temporal anchors like Drift and Multiple Spiky.

- **Explicit Semantic Analysis (ESA):** From this semantic representation[14] [31, 32], we devised an attribute, esa(k), which models text by means of its top-k closest related Wikipedia concepts (k = 1...10). Put differently, we made allowances for k distinct vectors, where each of them considers the k most semantically related Wikipedia concepts. This feature set theorizes that some temporally-anchored questions share the same array of underlying explicit topics. This might happens, for example, to questions regarding the different Jewish feasts.

- **WordNet (WN)/Collocations (Col):** WordNet[15] was used for checking semantic connections between pairs of terms in conformity to twenty-eight types including hypernyms and hyponyms. Thus, we interpreted as features one BoW representation per relation type, and its respective size. The most frequent sort of relation was also perceived as property. Analogously, we benefited from the eight kinds of collocations provided by

[13] For this purpose, we benefited from Mogura HPSG parser. Available at www.nactem.ac.uk/tsujii/enju/

[14] ticcky.github.io/esalib/

[15] wordnet.princeton.edu/

Oxford Dictionary[16]. This property set aims at modeling the notion that some terms have high probabilities of signaling an event when they are embodied in a specific WordNet class[35], and that some of these events might have high chances of being anchored temporally.

- **Predicate Analysis (PA):** We benefited from MontyLingua[17] for conducting predication. From this view, we generate bags of recognized subjects and verbs as well as arguments. In addition, we utilized the amount of detected predicates and the size of the bags. We further considered the highest frequent subject, verb and argument as attributes. Since the predicates outputted by Montylingua are n-ary relations, we expect that some of their components will indicate temporal anchors similarly to constituent parsing.

- **Misc:** Some extra characteristics include: a) the number of words in the longest, average and shortest sentences; b) the highest, average and lowest sentiment value in a sentence; c) the number of very positive, positive, neutral, negative and very negative sentences; and d) the number of words bearing of these five sentiment levels.

3.4. Models

In this work, we tried two approaches, one related to transfer learning ensemble (viz. Category-based Transfer Learning - CbTL Ensemble) and another one related to multi-view learning (viz. Dual Sequential Forward Floating Search - Dual SFFS). Although both strategies are aimed at boosting the prediction rate, they are radically different in spirit. In our empirical settings, both were tested in combination with several multi-class supervised classifiers of the following kinds:

- **Support Vector Machines (SVMs):** Non-probabilistic linear classifiers aimed at separating categories by a gap that is as large as possible. We benefited from the multi-core implementation supplied by Liblinear[18] [16, 40]. More specifically, we capitalized on two learners that our pre-liminar experiments showed to be most promising: L1-regularized L2-loss support vector classification (L1R/L2LOSS) and dual L2-regularized logistic regression (L2R/LR DUAL).

- **Bayes:** Probabilistic classifiers based on the theorem of Bayes with a strong independence assumption between the features. We profited from the multinomial and Bernoulli implementations supplied by OpenPR[19] [42]. Both combined with a traditional Laplace Smoothing.

- **Maximum Entropy Models (MaxEnt):** Probabilistic classifiers belonging to the family of exponential models. Particularly, MaxEnt does not assume that the features are conditionally independent [2]. In this work, we profited from an implementation mixed with L1 regularization[20]. These models have previously shown to be effective for similar classification tasks [27, 26].

- **Online learning:** Learning algorithms concerned with making decision with limited information [8]. We tested several approaches provided by Online Learning Library[21]: Log-Linear Models (SGD) [65], AROW [18], subgradient averaged hinge, several confidence weighted strategies [19, 23, 67, 68], and three passive aggressive methods [17].

CbTL Ensemble. The underlying idea behind this approach is determining which categories positively and negatively contribute to the recognition of temporal anchors across questions aiming at a particular target category. In other words, we conjecture that, in certain circumstances, some training material might be detrimental to the learning process and thus to the prediction of temporal anchors, and that this success/failure depends on the relationship between the target and training questions categories.

[16] oxforddictionary.so8848.com
[17] alumni.media.mit.edu/~hugo/montylingua/
[18] www.csie.ntu.edu.tw/ cjlin/libsvmtools/multicore-liblinear/
[19] www.openpr.org.cn/index.php/All/66-Naive-Bayes-EM-Algorithm/View-details.html
[20] www.nactem.ac.uk/tsuruoka/maxent/
[21] github.com/oiwah/classifier

More precisely, we hypothesize that some inferences can be negatively transferred from one category to the other, thus diminishing the overall performance of the system. Intuitively, for example, the word "*Christmas*" can be a strong indicator of periodicity if we are dealing with questions embodied in the category "*Food & Drink*", but much more weaker in the case of "*Society & Culture*". Therefore, harvesting questions from "*Food & Drink*" could be inappropriate to train models to deal with "*Society & Culture*", and the other way around.

As a natural consequence, this intuition suggests that distinct classifiers should be utilized for tackling different target inputs, more specifically it suggests building a classifier selection system (ensemble), in which each of the experts focuses on predicting the label of questions corresponding to a particular top-level category. Since all questions are categorized by the asker at posting time, i.e., assigned to an unique category, this kind of approach can be naturally applied to automatic question classification. Recall here that Yahoo! Answers question taxonomy system encompasses 26 distinct top-level question topics (e.g., Sports and Health), and accordingly, the proposed ensemble consists of 26 different experts.

In other words, our ensemble approach is a classifier selection system, where each of the 26 ensemble members are supposed to know well a part of the feature space and be responsible for objects in this part. In order to build each of these experts, we need to determine which category negatively affects the performance of another. In so doing, we designed a greedy algorithm that starts considering all data as training material, and systematically checks if there is a portion that hurts the performance by systematically removing all training data corresponding to each of the twenty-six Yahoo! Answers first level categories. For each of these automatically constructed subsets of data, we used SFFS [52] for determining its best array of attributes (see details in section 4). At the end of each iteration, it removes the data corresponding to the category that hurt the performance the most. If any, the algorithm stops.

In this way, CbTL Ensemble determines not only the relationship between training and testing data for each target category, but also its best working battery of attributes. In other words, from which categories the training material must be acquired as a means of enhancing the classification rate of a particular target question category, and the feature view derived thereof.

Dual SFFS. Multi-view learning has been integrated into both semi-supervised [10, 54, 63] and supervised learning methods [15, 25, 64]. Broadly speaking, approaches to build distinct views (e.g., redundant or collaborative) from a given feature set can be categorized into three groups: co-training, multiple kernel learning, and subspace learning [70]. Our approach falls into the last group as it constructs a latent subspace on top of two distinct collaborative views: one from the features harvested directly from the question itself (Φ_q), and the other considering any kind of property indirectly distilled from the question (Φ_{nq}). In this way, we aim at discovering which external and internal evidence must be gathered, and thus fused, in order to enhance the synergy between both sources, and as a natural consequence, to improve the recognition of the temporal anchors. Our approach generalizes the task of feature selection by inferring a latent subspace partitioning both feature spaces in such a way that these partitions work in tandem to enhance the system performance. Additionally, our method allows a feature selection algorithm to learn from the data the best relative contribution of these two disjoint views in the generated latent subspace.

In single-view learning, some algorithms generally search for a representative fixed-size set of characteristics as a means of singling out the most discriminative properties. However, other strategies do not impose this limit [12, 21, 22, 49, 59]. By and large, feature selection methods are categorized into three groups: filter, wrapper and embedded strategies (cf. [9, 11, 37]). In particular, wrapper techniques aim at finding a subset of features which produces the best classification rate according to the particularities of each classifier. Our approach uses a wrapper method that searches for two subsets $\phi_q \subseteq \Phi_q$ and $\phi_{nq} \subseteq \Phi_{nq}$ and their relative weight α so that the weighted linear combination of these two generated views brings about the highest classification rate, whilst taking advantage of the specific interactions between classifiers and datasets. That is to say, it constructs a latent layer that takes into account the synergy and relative importance between both sources of attributes.

More precisely, this latent layer is automatically constructed by adapting SFFS to this duality [52], which is outlined in algorithm 1. Unlike traditional SFFS, our Dual SFFS starts with an empty bag of attributes for each view ($\phi_q = \emptyset$ and $\phi_{nq} = \emptyset$), and at each iteration k , this procedure selects at most one property from each set of the available features (i.e., $\phi_q^k \in \Phi_q - \phi_q$ and $\phi_{nq}^k \in \Phi_{nq} - \phi_{nq}$). Thus, Dual SFFS can improve the classification rate by determining the best synergy of all linear combinations of the models produced when all potential selections of characteristics ϕ_q^k and ϕ_{nq}^k are added to ϕ_q and ϕ_{nq}, respectively. Note that, in some occasions, adding only one feature to one view brings about the best performance, meaning that only one ϕ_q^k or ϕ_{nq}^k, can be the empty set ($\emptyset$). After testing

all configurations, the best properties are definitively added to their specific view (i.e., ϕ_q or ϕ_{nq}), and the parameter α is updated accordingly. If both sets are empty, Dual SFFS finishes as no configuration enhanced the performance. Conversely, if any property was added, Dual SFFS starts what is called the backward step. This consists in checking if there is any nested attributes amongst the new sets of selected properties. In so doing, it removes each attribute and pair of properties (one from each view) chosen from iterations 1 to $k-1$. If any removal matches or improves the most the current best performance, its corresponding features are definitively removed from ϕ_q and ϕ_{nq} and put back into Φ_q and Φ_{nq}. The final outcome of Dual SFFS in contained in ϕ_q and ϕ_{nq} as well as the parameter α, which is the configuration of question and no-question traits (and their relative importance) that was found to have the best synergy.

Note that in order to linearly combine both views, a soft voting mechanism is computed so that each individual view produces a seven-dimensional vector regarded to as an estimate of the a-posteriori probability for each temporal

anchor. Soft voting tests several combined outputs by varying the parameter α from zero to 1 by a step of 0.01.

Algorithm 1: Dual SFFS

Input: Φ_q, Φ_{nq} (original feature spaces)
Result: Two features views ϕ_q and ϕ_{nq}
$\phi_q = \emptyset$;
$\phi_{nq} = \emptyset$;
$\alpha_{best} = 0$;
k=1;
bestPerformance=0;
repeat
  $\phi_q^{best@k} = \emptyset$;
  $\phi_{nq}^{best@k} = \emptyset$;
  $\alpha^{best@k} = 0.0$;
  forall $\phi_q^k \in \Phi_q - \phi_q \cup \emptyset$ **do**
    construct and test question view with $\phi_q \cup \phi_q^k$;
    forall $\phi_{nq}^k \in \Phi_{nq} - \phi_{nq} \cup \emptyset$ **do**
      construct and test no-question view with $\phi_{nq} \cup \phi_{nq}^k$;
      forall $\alpha = 0.0 \ldots 1$ ***step*** *0.01* **do**
        score = softVoting($view_q$, $view_{nq}$, α);
        if *score* > *bestPerformance* **then**
          $\phi_q^{best@k} = \phi_q^k$;
          $\phi_{nq}^{best@k} = \phi_{nq}^k$;
          $\alpha^{best@k} = \alpha$;
          bestPerformance=score;

  if $\phi_q^{best@k} \neq \emptyset$ *or* $\phi_{nq}^{best@k} \neq \emptyset$ **then**
    $\phi_q = \phi_q^{best@k} \cup \phi_q$;
    $\phi_{nq} = \phi_{nq}^{best@k} \cup \phi_{nq}$;
    $\phi_q^{bestr@k} = \emptyset$;
    $\phi_{nq}^{bestr@k} = \emptyset$;
    forall $\phi_q^k \in \phi_q - \phi_q^{best@k} \cup \emptyset$ **do**
      construct and test question view with $\phi_q - \phi_q^k$;
      forall $\phi_{nq}^k \in \phi_{nq} - \phi_{nq}^{best@k} \cup \emptyset$ **do**
        construct and test no-question view with $\phi_{nq} - \phi_{nq}^k$;
        forall $\alpha = 0.0 \ldots 1$ ***step*** *0.01* **do**
          score = softVoting($view_q$, $view_{nq}$, α);
          if *score* > *bestPerformance* **then**
            $\phi_q^{bestr@k} = \phi_q^k$;
            $\phi_{nq}^{bestr@k} = \phi_{nq}^k$;
            $\alpha^{best@k} = \alpha$;
            bestPerformance=score;

    $\alpha_{best} = \alpha^{best@k}$;
    $\phi_q = \phi_q - \phi_q^{bestr@k}$;
    $\phi_{nq} = \phi_{nq} - \phi_{nq}^{bestr@k}$;
  k++;
until $\phi_q^{best@k} = \emptyset$ *and* $\phi_{nq}^{best@k} = \emptyset$;

In terms of complexity, training a Dual SFFS model is much more demanding than learning a baseline model. For the sake of simplicity, let us assume that there is no effective removal during the backward step. As a rough approximation: we have at the first iteration, the baseline tests all its n features, and after each iteration it reduces its size by one during the forward step. Thus, after k iterations, the number of tests would be given by $k*n-k*(k-1)/2$. During the backward step, the baseline will perform $k-1$ tests, thus the number of backward trials at iteration k will be given by $(k-1)*(k-2)/2$. Combining the forward and backward steps, the baseline ends up performing

Learning Model	Baseline	Dual SFFS	CbTL Ensemble
Subgradient Averaged Hinge	0.7618	0.7655↑	0.7199
Confidence Weighted	0.7526	0.7532↑	0.7125
Soft Confidence Weighted	0.7505	0.7504	0.6941
AROW	0.7493	0.7564↑	0.7046
Passive Aggressive I	0.7489	0.7661↑	0.7237
Passive Aggressive II	0.7467	0.7590↑	0.7213
Soft Confidence Weighted II	0.7456	0.7429	0.7058
Bayes Multinomial	0.7432	**0.7721**↑	0.7431
Passive Aggressive	0.7374	0.7581↑	0.7044
MaxEnt	0.7270	0.7485↑	0.7177
Bayes Bernoulli	0.7213	0.7632↑	0.7189
LogLinear SGD	0.7196	0.7615↑	0.7431↑
Liblinear (L1R/L2LOSS)	0.5871	0.6593↑	0.5716
Liblinear (L2R/LR DUAL)	0.5423	0.6826↑	0.5639↑
Average (Std. Dev.)	0.7167 (± 0.066)	0.7457 (± 0.0329)	0.6944 (± 0.058)

Table 6: Results obtained by our two proposed models and the baseline, when combined with the different multi-class supervised learners. Results are expressed in MRR (test set), and the ↑ denotes an improvement wrt. the baseline system.

$k * n - (k-1)/2$ tests.

As for Dual SFFS, let us also assume that a feature was selected for each view at each iteration. Hence the number of forward tests is given by $n_1 * n_2, (n = n_1 + n_2)$ in the first iteration, therefore the amount of forward tests at the k iteration is given by $kn_1n_2 - k(k-1)(n_1+n_2)/2 + k(k-1)(2k-1)/6$. Regardless of the backward step, Dual SFFS performs at least $kn_1n_2 - k(k+1)(n_1+n_2)/2 + (k-1)(k(2k-1)-3)/6$ more trials than the baseline.

4. Experiments

In order to assess the performance of both proposed approaches, the experiments utilized the 6683 annotated questions obtained in section 3.2, which were randomly split into 4009 training (60%), 1337 testing (20%) and 1337 validation (20%) instances. Accordingly, held-out evaluations were conducted in all our experiments working on the same random splits. It is worth clarifying here that we utilize the test dataset for providing an unbiased evaluation of a final model fit on the training/evaluation datasets.

In all our experiments, a traditional SFFS algorithm was used for singling out the best array of features [52]. This process starts with an empty bag of properties and at each iteration it conducts a forward and a backward step. In the forward step, it adds the best performing feature, determined by testing each non-selected attribute together with all the properties in the bag. Thus the algorithm stops when no non-selected feature enhances the performance. Conversely, if any attribute was added to the bag, SFFS performs a backward step, where the algorithm checks the removal of each previously chosen feature contained in the bag. Ergo, the attribute corresponding to the largest growth in performance is removed and put back into the set of non-selected properties. The same happens to any removal that keeps the best performance (redundant/nested features). This backward phase is conducted iteratively until all removals diminish the performance.

We implemented a state-of-the-art **baseline** system, by capitalizing on SFFS and the high-dimensional feature set provided in section 3.3. In other words, we build effective traditional single-view models by checking the interactions of several features, while at the same time, benefiting from each learner mentioned in section 3.4.

Since all models output a confidence value for each candidate label, we took advantage of the Mean Reciprocal Rank (MRR) for assessing their performance. Basically, this metric is the multiplicative inverse of the position in the confidence ranking of the first correct label [66]. The MRR is then the average of the reciprocal ranks of the predictions obtained for a sample of questions.

4.1. Baseline

With regards to our best single-view model, our empirical outcomes point out to several interesting findings (see tables 6 and 7):

k	Type	Feature	MRR
1	web-snippet	BoW first three snippet bodies	0.7109
2	question-title	HPSG parser's lexical entries	0.7446
3	web-snippet	BoW top nine snippet titles	0.7558
4	question-body	HPSG parser's amount of different types	0.7674
5	question-title	Number of noun phrase clauses	0.7700
6	question-body	Lexicalised conj dependency relations	0.7717
7	question-body	Highest frequent nsubjpass lexical relation	0.7718
8	question-title	Lexicalised cc dependency relations	0.7719
9	question-body	Number of distinct aux relations	0.7729
10	question-title	Highest frequent nsubj lexical relation	0.7736
11	question-body	WordNet's Region Members found	0.7739
		Test set	0.7618

Table 7: Features integrated into the best baseline model (Subgradient Averaged Hinge).

1. A bird's eye view of the results points out to an average performance of 0.7166 (standard deviation of 0.066) across the different learners. In general, online learning strategies outperformed other kinds of learners (e.g., Bayes and MaxEnt), showing that Subgradient Averaged Hinge significantly improves the classification rate, reaping an MRR score of 0.7618. Noteworthily, this is a much less resource demanding learning algorithm in comparison to other tested approaches such as MaxEnt and Bayes. As displayed in Table 7, this algorithm also required only eleven characteristics to accomplish the highest prediction rate.
2. In detail, 71% of the performance (i.e., first three chosen features) achieved by testing several combinations of features were due to the titles and the description provided by web snippets together with the lexical entries found across the question title. This is relevant as it is expected that snippet titles are likely to contain question title words since these were used for the search. Note that a larger number of snippet titles were required in comparison to the number of snippet bodies. Needless to say, our results highlight that the first three web hits provides the most discriminative content. All in all, our results indicate that web search, i.e., insight mined from web snippets, is the most pertinent information to predict the temporal anchor of CQA question.
3. Additionally, noun phrase clauses (WHNP) together with two traits distilled from the lexicalised dependency tree view of the question title contributed to enhance the prediction rate. In particular, the highest frequent nominal subject (syntactic subject of a clause) across noun phrases. This feature is likely to signal the topical entity of the question, which can be the asker himself/herself.
4. As for question bodies, our empirical results also underscore the pertinence of syntactic subjects, but this time in passive form, harvested from the respective array of dependency trees. In the same spirit of the previous point, this characteristic reveals that askers express topical entities in the title using active voice, whereas the passive is used in descriptions.

All in all, the outcomes of our baseline emphasize subjects as key discriminative elements of temporal anchors integrated with on-line learning techniques. Our error analysis reveals that the three hardest categories to recognize were Multiple Bursty/Spiky, Permanent Non-recyclable and Periodic (see Table 8). As it relates to question categories (see Table 3), we discover on the test set that the MRR value widely ranges from 0.575 to 0.866, being *Health* the subject of the most successful performance, while the larger portion of errors were originated from the category *News & Events*. A similar picture is found in the validation set, the misclassification rate wages from 0.631 to 0.864, being *Science & Mathematics* the subject of the most successful predictions, whereas the larger fraction of miss-classifications came from *News & Events*. In effect, the Pearson Correlation Coefficient between both set of scores is 0.74, indicating a strong linear correlation. Other categories showing poor performance include *Dining Out* and *Environment*.

From another standpoint, Figure 1 reveals MRR achieved by questions grouped by the number of judges needed to set their class. This picture reveals that the performance substantially drops when the label was randomly chosen. This kind of questions was hard for both humans and automatic methods. We deem this as an effect of the multi-label nature of the temporal anchor of some questions. In fact, in about 30% of the cases, the best answer finished in the second position. Roughly speaking, the remaining groups achieve a similar performance, meaning that determining if a question is easy or hard to annotate by humans, it will not shed light into the difficulty for automatic models to

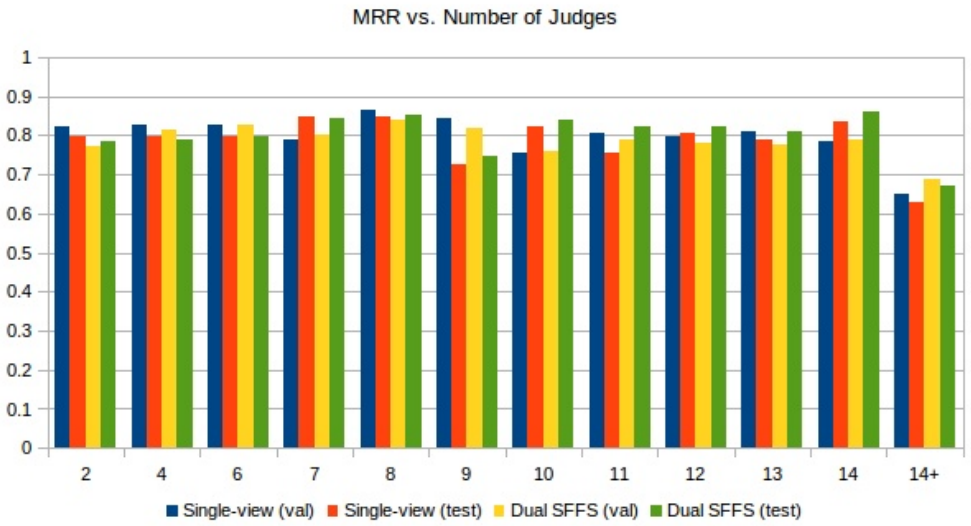

Figure 1: In the x-axis, the number of annotators required to set the temporal anchor, whereas the y-axis the MRR obtained by the best baseline/Dual SFFS model on the corresponding test/validation array of questions.

	Baseline				Dual SFFS			
	Validation Set		Test Set		Validation Set		Test Set	
Anchor	Accuracy	MRR	Accuracy	MRR	Accuracy	MRR	Accuracy	MRR
Drift	17.07%	0.4472	26.47%	0.5044	9.76%	0.2927	5.88%	0.2562
Multiple Spiky/Bursty	0%	0.2940	0%	0.1929	0%	0.1450	0%	0.1458
Other	63.51%	0.7925	59.08%	0.7647	69.96%	0.8444	65.27%	0.8212
Periodic	0.09%	0.3160	6.25%	0.3954	0%	0.1645	0%	0.1607
Permanent Non-Recyclable	3.16%	0.3627	0%	0.3352	9.47%	0.3998	3.41%	0.3634
Permanent Recyclable	80.33%	0.8907	77.89%	0.8806	77.33%	0.8826	74.80%	0.8682
Spiky/Bursty	34.94%	0.5758	33.33%	0.5242	39.76%	0.6009	34.67%	0.5659

Table 8: Outcomes achieved by the best baseline and Dual SFFS model wrt. each target temporal anchor.

predict its correct class. Along the same lines, the Pearson Correlation Coefficient between the average number of annotators (see Table 3) and the MRR achieved by each category is -0.136, indicating a weak anti-correlation, that is to say there is almost no relation between the difficulty for humans and the performance achieved per question category.

4.2. Dual SFFS

In all but two cases (i.e., both Soft Confidence Weighted methods), Dual SFFS improved the performance of the respective single-view model (see Table 6). In particular, the greater positive impact was observed in Support Vector methods (an increase of up to 25.83%), making these learners much more competitive to other approaches. On average, Dual SFFS reaps a score of 0.745, i.e., a growth of 4.05% and a decrease in the standard deviation. This means that dual-view strategies are much more independent upon the learning method. In effect, the best dual-view approach accomplishes an MRR value of 0.772, outperforming the best single-view model by 1.45%.

As a means of verifying the statistical significance between these top-two models, we bootstrap sampled their results twenty times and carried out a two-tailed t-test. Its outcome offers solid evidence in favour of a significant statistical difference between the top-two models ($p < 0.0001$). Ergo, in light of the best dual-view model (see Table 9), we can draw the following conclusions:

1. Interestingly enough, if we only consider the first three selected attributes of each view, Dual SFFS still outclasses the best single-view model (i.e., eleven properties). That is to say, a competitive performance was

achieved by building a simpler model.

2. The value of α points out to the important influence of both views in the prediction, being the question view more relevant than the other (i.e., 58% vs. 42%).
3. Thirteen out of the sixteen attributes were extracted from the question title (only three from the question body). This implies that discriminative characteristics are mainly found within the short context provided by the question title. Here, semantic/topical cues contributed the lion's share: words, amount of person names and indicators of tense. Note here that some dependency types can also give hints if the information conveyed has a temporal nature.
4. In juxtaposition, key elements across the bodies are time expressions, lists and the sentiment, in particular positive, of its sentences. We conjecture this is pertinent to recognize some opinionated questions.
5. All in all, features incorporated into the question view are radically different to the elements integrated into the best single view model. Curiously enough, our results support the thinking that there is necessity for enhancing the synergy between distinct feature sources. Still yet, our best models underscore that the NLP processing required to construct effective features include HPSG and dependency parsing as well as WordNet.
6. As for the non-question view, most discriminative attributes were distilled from the web instead of the CQA meta-data, even though, seven out of the fifteen properties were extracted from the asker self-description. We interpret this as the fact that community members express their main topic of interest in their profiles. For this reason, we find the BoW of root node harvested from dependency trees incorporated into the top-five features of this view. This feature can exploit the relationship between some topics and some temporal anchors, and the likelihood that community fellows are highly likely to prompt question on these topic of interests.
7. Another interesting finding emphasizes that snippet bodies were of less importance to the non-question view, contrary to the single-view model. More exactly, the top web features were extracted from their titles and URLs. The Internet archives also cooperated by counting the matching of the question month. We perceive this outcome as a results of the nature of hosts and URLs, that is to say some web-sites are linked to specific topics such as music and sports, while some URLs provide insight of temporal anchors, in particular new outlets. Note here that matching the month of question offered the best granularity.
8. With regard to the overlap between the best single-view model and the non-question view, we discover that snippet titles are key in both instances. Apart from that, both arrays of features are sharply different. In Dual SFFS, matching components of the question date becomes much more important than identifying some dependency relations.

In a nutshell, question and non-question properties proven to be pertinent, having question elements a greater influence on the final score (see table 1table:featuresDualView). Overall, effective single- and dual-view models are radically different, showing that each component view can underperform the best single-view model, but at the same time, their amalgamation accomplishes a higher classification rate. Broadly speaking, profiles and date hints become more relevant in a dual-view setting, while question bodies in a single-view one. Like our baseline system, the three hardest categories to predict were Multiple Bursty/Spiky, Permanent Non-recyclable and Periodic (see Table 8). As for question categories (see Table 3), we find out on the test set that the MRR value widely ranges from 0.627 to 0.890 corresponding to *News & Events* and *Health*, respectively. A similar picture is found in the validation set, the misclassification rate wages from 0.721 to 0.863, being *Society & Culture* the subject of the most successful predictions, whereas the larger fraction of miss-classifications came from *Arts & Humanities*. Interestingly enough, the Pearson Correlation Coefficient between both set of scores is -0.08, indicating that a linear correlation does not exist. Other categories showing poor performance include *Home & Garden* and *Yahoo! Products*.

From another angle, figure 1 unveils MRR accomplished by questions clustered by the amount of judges required to set their category. Like baseline models, the performance substantially decreases when the label was randomly chosen, but in the case of Dual SFFS, this drop in smaller. Roughly speaking, the remaining groups achieve a similar performance, meaning that determining if a question is easy or hard to annotate by humans, it will not shed light into the difficulty for automatic models to predict its correct class. Note also that Dual SFFS outclasses the single-view model in almost all cases where more six judges were needed. Along the same lines, the Pearson Correlation Coefficient between the average number of annotators (see Table 3) and the MRR achieved by each category is 0.012, indicating that a correlation does not exist.

	Question View			No-question View				
k	Type	Feature	MRR	Type	Feature	MRR	α	MRR
1	title	BoW without stop-words	0.7192	web-snippets	BoW first eight titles	0.7014	.58	0.7654
2	title	No. of person names	0.7309	web-snippets	First four hosts	0.7007	.62	0.7699
3	title	Highest frequent dependency type	0.7326	web-WBM	First seven snippets' month matches	0.6994	.60	0.7750
4	title	HPSG highest frequent tense	0.7305	web-snippets	First nine url's month matches	0.6992	.59	0.7773
5	title	No. of terms	0.7296	cqa-profiles	BoW (roots in lexicalised relations)	0.6993	.59	0.7791
6	title	BoW (punctuation)	0.7306	cqa-profiles	Highest frequent amod relation	0.7000	.59	0.7796
7	title	No. of Wh-adverb phrases	0.7296	web-WBM	First eight snippets' day matches	0.6995	.59	0.7809
8	body	BoW (time expressions)	0.7306	web-snippets	First four URLs' month matches	0.7003	.58	0.7819
8↑		-	0.7306	web-WBM	First eight snippets' day matches	0.6998	.58	0.7820
9	title	HPSG highest frequent voice	0.7300	web-snippets	First two snippets' day matches	0.7009	.58	0.7838
10	title	No. of WorNet's Part Holonyms found	0.7318	cqa-profiles	No. of adverbs	0.7014	.58	0.7845
11	title	Highest frequent iobj lexical relation	0.7316	cqa-profiles	BoW (adpositions)	0.7014	.58	0.7854
12	body	No. of List markers	0.7318	cqa-profiles	BoW (adjectives)	0.7014	.58	0.7857
13	title	No. of WorNet's Hyponyms found	0.7319	cqa-profiles	Lexicalised nummod dep. relations	0.7019	.58	0.7858
14	title	No. of Inverted declarative sentences	0.7309	web-WBM	First seven snippets' day matches	0.7016	.58	0.7860
15	body	No. of Very positive sentences	0.7318	web-snippets	First two url's year matches	0.7015	.58	0.7865
16	title	No. of adverbs	0.7316	cqa-profiles	Avg. minimum no. of children	0.7015	.58	0.7866
							Test set	0.7721

Table 9: Features integrated into the best Dual SFFS model (Bayes Multinomial). The ↑ denotes attributes removed after the backward step, while k the iteration and "WBM" stands for the Internet Archives.

4.3. Transfer Learning

Apart from two learners (see Table 6), the proposed transfer learning strategy worsens the results of our baseline, and it never defeats our Dual SFFS strategy. Anyway, by analyzing the outcomes outputted by the model achieving largest increase wrt. the baseline (LogLinear SGD), we discovered that the least portable category was *Travel*, which was removed when building four experts, that is to say when dealing with four distinct target categories. Conversely, training material coming from categories, such as *Pets*, *Social Science* and *Science & Mathematics*, was considered in all 26 cases.

Overall, our experiments suggest that our transfer learning ensemble was less effective due to the fact that most of the training material was necessary to build all the experts. In fact, results obtained by Dual SFFS ratify this finding as much more effective learning strategies could infer much more effective models by capitalizing on the whole material.

5. Conclusions

We have presented a new set of time-frame specific categories, which we obtained by fusing two distinct categories earlier developed by [50] and [29]. We have described the process and the results of a large crowdsourcing based human annotation effort of a question data set using up to fourteen workers. This effort resulted in a new corpus of 6683 English questions distilled form a very large data set crawled form Yahoo! Answers, labeled manually with the new time-frame specific categories.

Through a large number of experiments, we investigated the effectiveness of a wide variety of linguistic and web features compared to what was done in previous work. Using SFFS as baseline for multi-view learning, we observed that linguistic information is substantial for identification of temporal anchors, and that web search is substantial for identifying relevant text fragments. We showed that the use of a Dual version of SFFS improved the classification performance, but on different feature combinations compared to SFFS. We also introduced and explored the use of Category-based Transfer Learning (CbTL) ensembles in the context of CQA as an alternative to Dual SFFS, however, with less success as expected.

From a general point of view, we found out that humans and machines show different degree of difficulties when labeling questions from diverse topics. A topic that is easy to label by a human, might be difficult to label by a machine, and vice versa. Thus, at least in this task, the interpretability of machine decisions might be hard to achieve. Furthermore, our intuition that distinct classifiers should be utilized for different target inputs could not be verified by the results of our experiments using CbTL, since they were even lower than the results of SFFS.

We believe that the new high quality annotated question data set (publicly available at http://something.here.com) as well as our quantitative and qualitative data analyses provide a useful resource for future research in automatic question analysis, e.g., exploring alternative feature extraction strategies, machine learning algorithms or improving personalized adaptive search in CQA. We also believe that lifelong multi-label learning strategies seem to be key for temporal models.

6. 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, the German Federal Ministry of Education and Research (BMBF) through the project DEEPLEE (01IW17001) and the European Union's Horizon 2020 grant agreement No. 731724 (iREAD).

References

[1] Ittai Abraham, Omar Alonso, Vasilis Kandylas, Rajesh Patel, Steven Shelford, and Aleksandrs Slivkins. How many workers to ask?: Adaptive exploration for collecting high quality labels. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 473–482, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4069-4. doi: 10.1145/2911451.2911514. URL http://doi.acm.org/10.1145/2911451.2911514.

[2] Galen Andrew and Jianfeng Gao. Scalable training of l1-regularized log-linear models. In *Proceedings of the 24th international conference on Machine learning*, pages 33–40. ACM, 2007.

[3] Gabor Angeli and Jakob Uszkoreit. Language-independent discriminative parsing of temporal expressions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 83–92, 2013.

[4] Gabor Angeli, Christopher D Manning, and Daniel Jurafsky. Parsing time: Learning to interpret time expressions. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 446–455. Association for Computational Linguistics, 2012.

[5] Antoaneta Baltadzhieva and Grzegorz Chrupala. Question quality in community question answering forums: a survey. *SIGKDD Explorations*, 17:8–13, 2015.

[6] Steven Bethard. Cleartk-timeml: A minimalist approach to tempeval 2013. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 10–14, 2013.

[7] Jiang Bian, Yandong Liu, Eugene Agichtein, and Hongyuan Zha. Finding the right facts in the crowd: Factoid question answering over social media. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 467–476, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367561. URL http://doi.acm.org/10.1145/1367497.1367561.

[8] Avrim Blum. On-line algorithms in machine learning. In *In Proceedings of the Workshop on On-Line Algorithms, Dagstuhl*, pages 306–325. Springer, 1996.

[9] Avrim L. Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(12):245 – 271, 1997. ISSN 0004-3702. doi: http://dx.doi.org/10.1016/S0004-3702(97)00063-5. URL http://www.sciencedirect.com/science/article/pii/S0004370297000635. Relevance.

[10] Ulf Brefeld, Christoph Bscher, and Tobias Scheffer. Multi-view discriminative sequential learning. In Jo£o Gama, Rui Camacho, PavelB. Brazdil, AlpioMrio Jorge, and Lus Torgo, editors, *Machine Learning: ECML 2005*, volume 3720 of *Lecture Notes in Computer Science*, pages 60–71. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-29243-2. doi: 10.1007/11564096_11. URL http://dx.doi.org/10.1007/11564096_11.

[11] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16 – 28, 2014. ISSN 0045-7906. doi: http://dx.doi.org/10.1016/j.compeleceng.2013.11.024. URL http://www.sciencedirect.com/science/article/pii/S0045790613003066. 40th-year commemorative issue.

[12] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

[13] Angel X Chang and Christopher D Manning. Sutime: A library for recognizing and normalizing time expressions. In *Lrec*, volume 2012, pages 3735–3740, 2012.

[14] Minmin Chen, Kilian Q. Weinberger, and John Blitzer. Co-training for domain adaptation. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*, pages 2456–2464, 2011.

[15] Qiaona Chen and Shiliang Sun. Hierarchical multi-view fisher discriminant analysis. In *Proceedings of the 16th International Conference on Neural Information Processing: Part II*, ICONIP '09, pages 289–298, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-10682-8. doi: 10.1007/978-3-642-10684-2_32. URL http://dx.doi.org/10.1007/978-3-642-10684-2_32.

[16] Wei-Lin Chiang, Mu-Chu Lee, and Chih-Jen Lin. Parallel dual coordinate descent method for large-scale linear classification in multi-core environments. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1485–1494, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939826. URL http://doi.acm.org/10.1145/2939672.2939826.

[17] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7(Mar):551–585, 2006.

[18] Koby Crammer, Alex Kulesza, and Mark Dredze. Adaptive regularization of weight vectors. In *Advances in neural information processing systems*, pages 414–422, 2009.

[19] Koby Crammer, Mark Dredze, and Fernando Pereira. Confidence-weighted linear classification for text categorization. *J. Mach. Learn. Res.*, 13(1):1891–1926, June 2012. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2503308.2343704.

[20] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Comput. Surv.*, 51(1):7:1–7:40, January 2018. ISSN 0360-0300. doi: 10.1145/3148148. URL http://doi.acm.org/10.1145/3148148.

[21] Manoranjan Dash and Huan Liu. Feature selection for classification. *Intelligent data analysis*, 1(1):131–156, 1997.

[22] Nicoletta Dessì and Barbara Pes. Similarity of feature selection methods: An empirical study across data intensive classification tasks. *Expert Systems with Applications*, 42(10):4632–4642, 2015.

[23] Mark Dredze, Koby Crammer, and Fernando Pereira. Confidence-weighted linear classification. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 264–271, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390190. URL http://doi.acm.org/10.1145/1390156.1390190.

[24] Xin-yu Duan, Si-liang Tang, Sheng-yu Zhang, Yin Zhang, Zhou Zhao, Jian-ru Xue, Yue-ting Zhuang, and Fei Wu. Temporality-enhanced knowledgememory network for factoid question answering. *Frontiers of Information Technology & Electronic Engineering*, 19(1):104–115, Jan 2018. ISSN 2095-9230. doi: 10.1631/FITEE.1700788. URL https://doi.org/10.1631/FITEE.1700788.

[25] Jason D. R. Farquhar, David R. Hardoon, Hongying Meng, John S. Taylor, and Sándor Szedmák. Two view learning: SVM-2K, theory and practice. In *NIPS*, 2005.

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

[27] Alejandro Figueroa and John Atkinson. Ensembling classifiers for detecting user intentions behind web queries. *IEEE Internet Computing*, 20(2):8–16, Mar 2016. ISSN 1089-7801.

[28] Alejandro Figueroa and Günter Neumann. Category-specific models for ranking effective paraphrases in community question answering. *Expert Syst. Appl.*, 41(10):4730–4742, 2014.

[29] Alejandro Figueroa, Carlos Gómez-Pantoja, and Ignacio Herrera. Search clicks analysis for discovering temporally anchored questions in community question answering. *Expert Systems with Applications*, 50:89 – 99, 2016. ISSN 0957-4174. doi: http://dx.doi.org/10.1016/j.eswa.2015.12.016. URL http://www.sciencedirect.com/science/article/pii/S0957417415008180.

[30] Michele Filannino, Gavin Brown, and Goran Nenadic. Mantime: Temporal expression identification and normalization in the tempeval-3 challenge. *arXiv preprint arXiv:1304.7942*, 2013.

[31] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc. URL http://dl.acm.org/citation.cfm?id=1625275.1625535.

[32] Evgeniy Gabrilovich and Shaul Markovitch. Wikipedia-based semantic interpretation for natural language processing. *J. Artif. Int. Res.*, 34 (1):443–498, March 2009. ISSN 1076-9757. URL http://dl.acm.org/citation.cfm?id=1622716.1622728.

[33] Poonam Gupta and Vishal Gupta. Article: A survey of text question answering techniques. *International Journal of Computer Applications*, 53(4):1–8, September 2012. Full text available.

[34] Sanda M. Harabagiu and Cosmin Adrian Bejan. An answer bank for temporal inference. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006.*, pages 741–746, 2006.

[35] Yoonjae Jeong and Sung-Hyon Myaeng. Using wordnet hypernyms and dependency features for phrasal-level event recognition and type classification. In *European Conference on Information Retrieval*, pages 267–278. Springer, 2013.

[36] Hideo Joho, Adam Jatowt, and Roi Blanco. A survey of temporal web search experience. In *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web*, pages 1101–1108, 05 2013.

[37] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(12):273 – 324, 1997. ISSN 0004-3702. doi: http://dx.doi.org/10.1016/S0004-3702(97)00043-X. URL http://www.sciencedirect.com/science/article/pii/S000437029700043X. Relevance.

[38] Mirella Lapata and Alex Lascarides. Learning sentence-internal temporal relations. *Journal of Artificial Intelligence Research*, 27:85–117, 2006.

[39] Kenton Lee, Yoav Artzi, Jesse Dodge, and Luke Zettlemoyer. Context-dependent semantic parsing for time expressions. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1437–1447, 2014.

[40] Mu-Chu Lee, Wei-Lin Chiang, and Chih-Jen Lin. Fast matrix-vector multiplications for large-scale logistic regression on shared-memory systems. In *2015 IEEE International Conference on Data Mining*, pages 835–840, Nov 2015. doi: 10.1109/ICDM.2015.75.

[41] Damien Leprovost, Abrouk Lylia, and David Gross-Amblard. Discovering implicit communities in web forums through ontologies. 10: 93–103, 01 2012.

[42] David D. Lewis. *Naive (Bayes) at forty: The independence assumption in information retrieval*, pages 4–15. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998. ISBN 978-3-540-69781-7. doi: 10.1007/BFb0026666. URL http://dx.doi.org/10.1007/BFb0026666.

[43] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014. URL http://www.aclweb.org/anthology/P/P14/P14-5010.

[44] Zide Meng, Fabien Gandon, and Catherine Faron Zucker. Overlapping community detection and temporal analysis on q&a sites. In *Web Intelligence*, volume 15, pages 115–142. IOS Press, 2017.

[45] Sewon Min, Min Joon Seo, and Hannaneh Hajishirzi. Question answering through transfer learning from large fine-grained supervision data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 510–517, 2017.

[46] Amit Mishra and Sanjay Kumar Jain. A survey on question answering systems with classification. *J. King Saud Univ. Comput. Inf. Sci.*, 28(3): 345–361, July 2016. ISSN 1319-1578. doi: 10.1016/j.jksuci.2014.10.007. URL https://doi.org/10.1016/j.jksuci.2014.10.007.

[47] Yusuke Miyao, Takashi Ninomiya, and Jun'ichi Tsujii. *Corpus-Oriented Grammar Development for Acquiring a Head-Driven Phrase Structure Grammar from the Penn Treebank*, pages 684–693. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005. ISBN 978-3-540-30211-7. doi: 10.1007/978-3-540-30211-7_72. URL https://doi.org/10.1007/978-3-540-30211-7_72.

[48] Günter Neumann and Bogdan Sacaleanu. Dfki's lt-lab at the CLEF 2005 multiple language question answering track. In *Working Notes for CLEF 2005 Workshop co-located with the 9th European Conference on Digital Libraries (ECDL 2005), Wien, Austria, September 21-22, 2005.*, 2005.

[49] Jana Novoviov, Petr Somol, and Pavel Pudil. Oscillating feature subset search algorithm for text categorization. In JosFrancisco Martnez-Trinidad, JessAriel Carrasco Ochoa, and Josef Kittler, editors, *Progress in Pattern Recognition, Image Analysis and Applications*, volume 4225 of *Lecture Notes in Computer Science*, pages 578–587. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-46556-0. doi: 10.1007/11892755_60. URL http://dx.doi.org/10.1007/11892755_60.

[50] Aditya Pal, James Margatan, and Joseph A. Konstan. Question temporality: identification and uses. In *CSCW '12 Computer Supported Cooperative Work, Seattle, WA, USA, February 11-15, 2012*, pages 257–260, 2012.

[51] Barun Patra. A survey of Community Question Answering. *ArXiv e-prints*, May 2017.

[52] Pavel Pudil, Jana Novovicová, and Josef Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119 – 1125, 1994. ISSN 0167-8655. doi: https://doi.org/10.1016/0167-8655(94)90127-9. URL http://www.sciencedirect.com/science/article/pii/0167865594901279.

[53] Salvatore Romeo, Giovanni Da San Martino, Alberto Barrón-Cedeño, Alessandro Moschitti, Yonatan Belinkov, Wei-Ning Hsu, Yu Zhang, Mitra Mohtarami, and James R. Glass. Neural attention for learning to rank questions in community question answering. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 1734–1745, 2016.

[54] David S Rosenberg, Vikas Sindhwani, Peter L Bartlett, and Partha Niyogi. Multiview point cloud kernels for semisupervised learning [lecture notes]. *Signal Processing Magazine, IEEE*, 26(5):145–150, 2009.

[55] Estela Saquete, Patricio Martínez-Barco, Rafael Muñoz, and José Luis Vicedo González. Splitting complex temporal questions for question answering systems. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain.*, pages 566–573, 2004.

[56] Yikang Shen, Wenge Rong, Zhiwei Sun, Yuanxin Ouyang, and Zhang Xiong. Question/answer matching for CQA system via combining lexical and sequential information. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 275–281, 2015.

[57] Shashank Shrivastava, Mitesh Khapra, and Sutanu Chakraborti. A concept driven graph based approach for estimating the focus time of a document. In *Mining Intelligence and Knowledge Exploration - 5th International Conference, MIKE 2017, Hyderabad, India, December 13-15, 2017, Proceedings*, pages 250–260, 2017.

[58] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 254–263, 2008.

[59] Petr Somol, J Novovičová, P Pudil, et al. Dynamic oscillating search algorithm for feature selection. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

[60] Ivan Srba and Mária Bieliková. A comprehensive survey and classification of approaches for community question answering. *TWEB*, 10: 18:1–18:63, 2016.

[61] Ilija Subasic and Carlos Castillo. The effects of query bursts on web search. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010, Toronto, Canada, August 31 - September 3, 2010, Main Conference Proceedings*, pages 374–381, 2010.

[62] Shiliang Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7-8):2031–2038, 2013. doi: 10.1007/s00521-013-1362-6. URL https://doi.org/10.1007/s00521-013-1362-6.

[63] Shiliang Sun and John Shawe-Taylor. Sparse semi-supervised learning using conjugate functions. *The Journal of Machine Learning Research*, 11:2423–2455, 2010.

[64] Sandor Szedmak and John Shawe-Taylor. Synthesis of maximum margin and multiview learning using unlabeled data. *Neurocomput.*, 70 (7-9):1254–1264, March 2007. ISSN 0925-2312. doi: 10.1016/j.neucom.2006.11.012. URL http://dx.doi.org/10.1016/j.neucom.2006.11.012.

[65] Yoshimasa Tsuruoka, Jun'ichi Tsujii, and Sophia 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*, pages 477–485. Association for Computational Linguistics, 2009.

[66] Ellen M Voorhees et al. The trec-8 question answering track report. In *TREC*, volume 99, pages 77–82, 1999.

[67] Jialei Wang, Peilin Zhao, and Steven C. H. Hoi. Exact soft confidence-weighted learning. *CoRR*, abs/1206.4612, 2012. URL http://arxiv.org/abs/1206.4612.

[68] Jialei Wang, Peilin Zhao, and Steven C. H. Hoi. Soft confidence-weighted learning. *ACM Trans. Intell. Syst. Technol.*, 8(1):15:1–15:32, September 2016. ISSN 2157-6904. doi: 10.1145/2932193. URL http://doi.acm.org/10.1145/2932193.

[69] Fei Wu, Xinyu Duan, Jun Xiao, Zhou Zhao, Siliang Tang, Yin Zhang, and Yueting Zhuang. Temporal interaction and causal influence in community-based question answering. *IEEE Trans. Knowl. Data Eng.*, 29(10):2304–2317, 2017. doi: 10.1109/TKDE.2017.2720737. URL https://doi.org/10.1109/TKDE.2017.2720737.

[70] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *CoRR*, abs/1304.5634, 2013. URL http://arxiv.org/abs/1304.5634.

[71] Zhijie Xu and Shiliang Sun. An algorithm on multi-view adaboost. In *Neural Information Processing. Theory and Algorithms - 17th International Conference, ICONIP 2010, Sydney, Australia, November 22-25, 2010, Proceedings, Part I*, pages 355–362, 2010.

[72] Sha Yuan, Yu Zhang, Jie Tang, and Juan Bautista Cabotà. Expert finding in community question answering: A review. *CoRR*, abs/1804.07958, 2018. URL http://arxiv.org/abs/1804.07958.

[73] Xiaoshi Zhong and Erik Cambria. Time expression recognition using a constituent-based tagging scheme. In *Proceedings of the 2018 World*

Wide Web Conference on World Wide Web, pages 983–992. International World Wide Web Conferences Steering Committee, 2018.

[74] Xiaoshi Zhong, Aixin Sun, and Erik Cambria. Time expression analysis and recognition using syntactic token types and general heuristic rules. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 420–429, 2017.