Context-Dependent Conceptualization

Dongwoo Kim[*][†], **Haixun Wang**[‡], **Alice Oh**[†]
[†] Computer Science Department, KAIST, Daejeon, Korea
[‡] Microsoft Research Asia, Beijing, China
dw.kim@kaist.ac.kr, haixunw@microsoft.com, alice.oh@kaist.edu

Abstract

Conceptualization seeks to map a short text (i.e., a word or a phrase) to a set of concepts as a mechanism of understanding text. Most of prior research in conceptualization uses human-crafted knowledge bases that map instances to concepts. Such approaches to conceptualization have the limitation that the mappings are not context sensitive. To overcome this limitation, we propose a framework in which we harness the power of a probabilistic topic model which inherently captures the semantic relations between words. By combining latent Dirichlet allocation, a widely used topic model with Probase, a large-scale probabilistic knowledge base, we develop a corpus-based framework for context-dependent conceptualization. Through this simple but powerful framework, we improve conceptualization and enable a wide range of applications that rely on semantic understanding of short texts, including frame element prediction, word similarity in context, ad-query similarity, and query similarity.

1 Introduction

Mapping a short text (i.e., a word or a phrase) to concepts is an important problem in natural language processing. For instance, the word *apple* can be mapped to the concepts *fruit* and *food* or *company* and *firm*, and a phrase *apple orchard* can be mapped to a *piece of land* with *fruit trees*. This mapping of words to their most appropriate concepts is what we seek to accomplish in conceptualization. It is an important component of semantic understanding tasks, for example, matching of a search query to an advertisement. The query-ad matching task depends heavily on conceptualization because more full-blown analysis such as syntactic parsing is often not applicable nor helpful for understanding queries or ad keywords. To solve this conceptualization problem, we can use knowledge bases that explicitly represent the word-to-concept relationships, but knowledge bases alone often lack coverage and context-sensitivity for this task. In this paper, we propose a wide-coverage context-dependent solution for conceptualization. Our solution discovers the semantic context of words with a probabilistic topic model and maps the words to the most appropriate concepts in a large-scale probabilistic knowledge base containing millions of concept-instance mappings.

An unsolved challenge in conceptualization is the context-sensitive nature of word-to-concept mappings, for instance how the concept of the word *apple* changes from *fruit* and *food* when used together with *orchard* to *company, firm* and *market leader* when used together with *iPad*. A promising solution for this problem uses a large-scale probabilistic knowledge base [Wu *et al.*, 2012] which can disambiguate the concept of *apple* when used together with *Microsoft*, as they both belong to the concept of *company*. However, for many cases, semantically related words are not related in the concept space. An example is the pair of words *iPad* and *apple*, which are semantically related, but are tied to concepts *device, product, tablet* and *company, firm, fruit*, respectively. Hence, the challenge left unresolved is that the knowledge base does not capture the semantic relationships among words.

We approach this problem with a simple two-stage solution combining a topic model and a probabilistic knowledge base such that we can consider both semantic relationships among words for modeling the context, and conceptual relationships between words and concepts for mapping the words to concepts within the given context. A probabilistic topic model [Blei *et al.*, 2003], which estimates how words are semantically related based on their general co-occurrence statistics, is a natural candidate for capturing the semantic relationships. And a probabilistic knowledge base, Probase [Wu *et al.*, 2012], which models the probabilistic concept-instance mappings, is a good resource for capturing the conceptual relationships. We propose a two-stage approach in which we first estimate the topical context using LDA and then estimate the most likely concepts given the topic context using Probase.

The rest of this paper is organized as follows. In Section 2, we describe background research in conceptualization and topic modeling. In Section 3, we explain the details of context-dependent conceptualization (CDC) and the results of conceptualization experiments. We show how CDC

[*] This research was done while the first author was an intern at Microsoft Research Asia.

outperforms a previous method on the experiments of frame element conceptualization and word similarity in context. In Section 4, we describe sentence-level conceptualization, which aims to combine concepts of multiple instances in a sentence. We evaluate sentence-level conceptualization on tasks of ad-query and query-URL similarity. In Section 5, we conclude the paper with discussions of CDC and future directions.

2 Background

The main contribution of this work is in proposing a simple but effective framework for combining a probabilistic topic model and a large-scale probabilistic knowledge base to solve the problem of context-dependent conceptualization. In this section, we describe the two tools we use, Probase, containing millions of probabilistic concept-instance relationships, and LDA, a probabilistic topic model to capture the semantic relationships among words and phrases.

2.1 Probase

Probase is a probabilistic knowledge base consisting of more than three million concepts automatically extracted using syntactic patterns (such as the Hearst patterns) from billions of Web pages [Wu *et al.*, 2012]. A unique advantage of Probase is that the concept-instance relationships are probabilistic such that for each concept c and each instance word w, Probase specifies the probability of each instance belonging to that concept, $P(w|c)$. It also specifies the probabilities of the reverse direction, $P(c|w)$. This probabilistic nature of Probase allows us to compute and combine the probabilities of concepts when there are two or more instance words in a phrase or sentence.

2.2 Conceptualization using Probase

Conceptualization is useful for many IR and NLP tasks such as query understanding [Hua *et al.*, 2013], web table understanding [Wang *et al.*, 2012], and sentiment analysis [Cambria *et al.*, 2013]. Conceptualization of a single word given a knowledge base like Probase is simply a look-up process, with the output of a set of possible concepts with probabilities for each concept. Given a short text, such as a phrase or a sentence consisting of two or more nouns, each of which can be mapped to a set of concepts with probabilities, we need a way to combine the multiple sets of possible concepts. A previous conceptualization technique [Song *et al.*, 2011] uses a simple probabilistic framework to combine the probabilities of the concepts from the two or more words. It estimates the posterior probability of concepts given a short text by using a naïve Bayes approach in which the common concepts of several instances get high posterior probabilities over the concept space. For example, given a short text *iPhone and Windows phone*, this can be parsed into two instances, *iPhone* and *Windows phone*. The posterior shapes a high probability over the concepts that are shared by both instances, in this case *smart phone*, and *mobile phone*.

The limitation of this previous conceptualization technique is that when it tries to conceptualize a short text for which the instances do not share any concepts, such as *iPad and apple*, it would assign high probabilities for the concept *food* as well as the concepts *device* and *company*. This is counterintuitive because we think of the words *iPad* and *apple* to be closely related, but they are only related in the semantic space, not in the concept space. Therefore, we need a way to consider both the concept relationships and the semantic relationships.

2.3 Topic Modeling

One intuitive way to model semantic relationships among instances is by using a probabilistic topic model, such as latent Dirichlet allocation (LDA) [Blei *et al.*, 2003]. LDA discovers a latent pattern of semantic themes, called *topics*, by computing and maximizing the posterior distribution of the documents which are assumed to be generated from those latent topics. The topics that LDA discovers are multinomial distributions over the vocabulary, and because they indirectly represent the co-occurrence patterns of the words in the vocabulary, they essentially capture the semantic relationships of the words.

Two related papers also introduce incorporating a knowledge base and a probabilistic topic model. In [Boyd-Graber *et al.*, 2007], LDA-WordNet (LDAWN) embeds the WordNet ontology into LDA for an improvement in word sense disambiguation. Our work differs from LDAWN in that our goal is general word- and sentence-level conceptualization, which enables more variety of semantic tasks than word sense disambiguation. In [Chemudugunta *et al.*, 2008], the concept-topic model (CTM) uses a small set of human-crafted ontology to infer the probability of instances given concepts, $p(w|c)$. Our work differs from CTM in that we use a large-scale probabilistic knowledge base which directly defines $p(w|c)$ to learn the context sensitive concept distribution over the given sentences and phrases.

3 Context-Dependent Conceptualization

We describe our method for improving word conceptualization with the help of surrounding text as context. We show that our context-dependent conceptualization (CDC) outperforms previous methods in two experiments, predicting an unseen frame element and measuring word similarity in context.

3.1 Using Topics to Improve Conceptualization

The intuition behind CDC is that the topic distribution of a short text serves as a guide for conceptualizing each instance word. With our running example, CDC would work as follows:

- Given a short text ‘apple and iPad’, LDA assigns a high probability to a topic related to computer companies.
- For the same text, Probase assigns high probabilities to mapping the instance word ‘apple’ to concepts *fruit, firm*.
- Within the company related topic, *firm* has a high probability, while *fruit* has a low probability.
- Computing the weighted sum of the probability of topic given the text and the probability of concept given the topic, we can determine the concept *firm* is more probable than *fruit* for the instance term ‘apple’.

Table 1: Top ten concepts of context-dependent conceptualization (CDC) for *apple* and *jordan*. The original concept vector of *apple* contains fruit, company, food concepts, but CDC separates the company-related concepts and food-related concepts depending on context. CDC also separates the concepts of *jordan* to the country and basketball player in two different contexts.

Concept of *apple*			Concept of *jordan*		
apple	apple and orchard	apple and ipad	jordan	jordan and basketball	jordan and iraq
fruit	fruit	company	country	player	country
company	food	client	arab country	team	state
food	tree	tree	state	state	arab country
fresh fruit	plant	corporation	place	professional athlete	arab state
fruit tree	crop	computer	nation	great player	muslim country
brand	fruit tree	software company	arab state	offensive force	arab nation
crop	wood	oems	muslim country	nike shoe	islamic country
flavor	juice	laptop	others	wing player	middle eastern country
item	flavor	personal computer	middle eastern country	signature shoe	arab government
manufacturer	firm	host	case	good player	regime

The challenge and contribution of this work is in formalizing the above steps with LDA and Probase. We describe the details below.

Estimating Topic Distributions

In the first step of CDC, we infer the topic distribution of a given phrase with a topic model. We first train LDA with an external corpus using collapsed Gibbs sampling, and we obtain from it the sample assignments C_{wk}, the number of times term w is assigned to topic k. We will describe the details of how we trained the model in the section "Training a Topic Model".

Given a trained model C, we can infer the topics of the words in a sentence via a streaming sampling method [Yao *et al.*, 2009]. Let $\vec{s}$ be the sequence of word indices of a target sentence, and $\vec{z}$ be the topic assignment vector of sentence $\vec{s}$. We infer the posterior of $\vec{z}$ using collapsed Gibbs sampling based on the trained model as follows:

$$\begin{aligned} & p(z_i = k|\vec{s}, z_{-i}, C) \\ & \quad \propto (n_{\cdot k} + \alpha) \times \frac{C_{s_i k} + n_{s_i k} + \beta}{\sum_w C_{wk} + n_{wk} + |W|\beta}, \end{aligned} \tag{1}$$

where n_{wk} is the number of times term w is assigned to topic k of sentence s, $|W|$ is the size of the vocabulary, and α and β are hyper-parameters for document-topic and word-topic distributions, respectively. We use the dot notation to summarize the index, and z_{-i} to denote the topic assignments except word i in the sentence. Finally, we estimate the posterior topic probability, $p(z_i)$, of each word s_i in the sentence through the sampling results.

Estimating Concept Distributions

In the second step of CDC, we estimate the concept distribution for a sentence by computing the probability of each concept based on the topic distribution of the sentence. Probase has two distinct vocabularies, namely the instance vocabulary I and the concept vocabulary C, but LDA does not distinguish the type of words and uses the union of both vocabularies as a word vocabulary (i.e. $W = C \cup I$). By including the instance terms and the concept terms in the same vocabulary, LDA may discover a topic with high probabilities for words that are semantically related but conceptually distant, such as *iPhone* and *computers*. Formally, we compute the probability of concept c given instance w with its context topics as follows:

$$p(c|w, z) \propto p(c|w) \sum_k \pi_{wk} \phi_{ck}, \tag{2}$$

$$\phi_{ck} = \frac{C_{ck} + \beta}{\sum_w C_{wk} + |W|\beta},$$

where c is the index of the concept-term in the vocabulary, $\pi_{wk} = p(z_w = k)$ is the inferred topic distribution, $p(c|w)$ is defined in Probase, ϕ_{ck} is the probability of concept c given topic k. The summation term regularizes the probability of concept c given the context of instance w by considering the topic distribution π_{wk}.

Table 1 shows the context-dependent conceptualization result of the instance term *apple* in the context of 'apple and iPad'. The original concepts of apple include *fruit, company*, and *food*, but the context-dependent conceptualization filtered out the fruit-related concepts.

Training a Topic Model

LDA discovers topics in a purely corpus-driven way. Therefore, to train a set of topics with a broad coverage of most of the important concepts in Probase, we fit LDA with a corpus of about 3 million Wikipedia documents. To estimate the posterior, we use collapsed Gibbs sampling [Griffiths and Steyvers, 2004], with 1,000 iterations and 500 burn-in samples. We set α, the document topic proportion hyper-parameter, as 0.1 and β, the word-topic distribution hyper-parameter as 0.01. We set k, the number of topics, to be $\{100, 200, 300\}$, for three sets of topics for various experiments of context-dependent conceptualization.

3.2 Experiment 1 : Frame Element Conceptualization

Semantic role labeling (SRL) is an NLP task which detects the semantics associated with a verb or a predicate of a sentence and classifies the elements of the sentence into specific roles. Several methods including supervised learning [Gildea and Jurafsky, 2002] and semi-supervised learning [Fürstenau and Lapata, 2009] have been suggested to solve this problem.

FrameNet [Baker *et al.*, 1998] is a valuable lexical database containing more than 170,000 sentences manually annotated for semantic roles. Each sentence is annotated as a specific frame, where a frame represents the overall semantics of the sentence with corresponding predicates, each frame has its own frame elements (FE) as arguments. For example, given a sentence *"the boys grill their catches on an open fire"*, this sentence corresponds to the *Apply_heat* frame with three FEs. *the boys* corresponds to the *cook* FE, *their catches* corresponds to the *food* FE, and *an open fire* corresponds to the *heating_instrument* FE. While the careful annotations of FrameNet make it a useful resource for supervised and unsupervised learning of SRL, it would be a more complete resource if the size and coverage could be expanded. One can try to expand the instances of FEs by using a language model (LM) or a topic model based on the sentences of FrameNet. The LM approach does not fit the problem, as it would be based on the observed documents, therefore instances of FEs which do not exist in the training set cannot have high posterior probabilities. A topic model based expansion also would have a problem because topics only represent semantic relatedness of the terms, as we can see in the example of *iPad* and *apple*, in which case their semantic relatedness would not be useful to find new instances of FE.

Expanding the coverage of FrameNet requires a method to predict unseen instances of the FEs, and those instances must be conceptually and semantically similar to the observed instances of the FEs. CDC is one solution to expand the coverage of FEs by finding new words that are similar to the observed instances. For instance, we can easily imagine that the *food* FE corresponds to concepts such as *food*, and *fruit*, and based on these concepts we can obtain instances such as *orange*, *rice*, and *bread*.

For the experiment, we first collect all sentences that contain a specific FE and parse the sentence to find multi-word expressions [Song *et al.*, 2011]. We conceptualize each instance of the FE within the sentence and take an average of the concept vectors as concepts of FE. The first two columns of Table 2 show the top concepts of the *Heat_source* FE, and we can see the *heat source* concept has a high probability as expected. Based on the concept probability, we further compute the probability of each instance as follows :

$$p(w|\text{FE}) \propto \sum_{c} p(w|c)p(c|\text{FE}), \tag{3}$$

where $p(w|c)$ is defined in Probase, and $p(c|\text{FE})$ can be computed by Equation 2. We list the top instances of *Heat_source* FE in the right two columns of Table 2. Several different sources of heating instrument such as stove, lamp, and candle are found by this method, and half of the top ten instances are not seen in the training samples.

To measure the quality of instances expanded by CDC, we use five-fold cross validation on the FrameNet dataset. We compute the predictive probability of instances for each FE from training set and compute the test set likelihood of FE instances by using the trained result. We compare this result with the naive approach in which we use Probase original concept vectors of FE instances for conceptualization. Table 3 shows per-word heldout log-likelihood of each fold for the naive approach and CDC with 100, 200, and 300 topics. CDC generates better likelihood than naive approach, and as the number of topics increases, the predictive likelihood increases.

Table 2: Top concepts and predicted instances of *Heat_source* FE. The predicted instance list contains several different heat sources, such as stove and hair dryer. Instances with $*$ are not seen in the training set.

Concept	$p(c\|\text{FE})$	**Instance**	$p(w\|\text{FE})$
heat source	0.19	stove	0.00019
place	0.17	radiator $*$	0.00015
accessory	0.09	oven	0.00015
large part	0.07	grill $*$	0.00014
large metal	0.04	heater $*$	0.00013
essential	0.03	fireplace $*$	0.00013
supplement	0.03	car $*$	0.00013
heat	0.03	lamp $*$	0.00013
kitchen appliance	0.02	hair dryer $*$	0.00012
compound	0.02	candle $*$	0.00012

Table 3: Per-word Heldout log-likelihood of frame elements with five-fold validation. Our approach (CDC-#topic) produces better predictive likelihood than a naive conceptualization approach based on Probase without context.

Likelihood	Naive	CDC-100	CDC-200	CDC-300
Fold 1	-4.716	-3.401	-3.385	**-3.378**
Fold 2	-4.728	-3.409	-3.393	**-3.389**
Fold 3	-4.741	-3.432	-3.417	**-3.410**
Fold 4	-4.727	-3.413	-3.399	**-3.392**
Fold 5	-4.740	-3.433	-3.417	**-3.413**

3.3 Experiment 2 : Context-Dependent Word Similarity

Our second experiment applies CDC to measure the similarity between words presented in sentences. A common dataset for word similarity, WordSim-353 [Finkelstein *et al.*, 2001] consists of pairs of words without context. However, as we saw in the *apple* example, homonymous and polysemous words vary their meaning depending on the context. Recent research [Huang *et al.*, 2012] presented a new dataset to measure the similarity between words in context, where the dataset provides two words within sentences, and ten non-experts annotated their similarity scores.

> ...the lightduty Ridgeline, won Truck of the Year from "Motor Trend " magazine in 2006 (also in 2006, the redesigned Civic won **Car** of the Year from the magazine, giving Honda a rare double win of Motor Trend honors). Mountain bikes ...

> ...Tamil Nadu has seen major investments in the **automobile** industry over many decades manufacturing cars, railway coaches, battle-tanks, tractors, motorcycles, automobile spare parts and accessories, tyres and heavy vehicles ...

The above two excerpts are from the dataset, and they show the words **Car** and **automobile** in sentential contexts. After reading each pair of excerpts, annotators judge the similar-

Table 4: Word similarity in context.

Model	Correlation
BOW-window2	0.31
BOW-window5	0.30
Naive concept vector	0.44
CDC-Topic 100	**0.52**
CDC-Topic 200	0.50
CDC-Topic 300	0.50

ity of highlighted words on a zero-to-ten scale. For example, in the above excerpts, annotators rate the similarity between **Car** and **automobile**, and the average similarity was 8.8 among ten annotators. We conceptualize the highlighted words using CDC and measure the cosine similarity between the concept distributions of the two words, and for the above example, CDC scores 0.71.

For evaluation, we compute Pearson's correlation between our model-based similarities and human judgment. To measure the similarity between two words, we use cosine similarity using the concept vectors of the words. We compare CDC against two different baselines, a bag-of-words (BOW) approach and a naive concept vector approach as in experiment 1. For BOW, we construct a bag-of-words within a k-word window before and after the target word, with k of two and five, such that a BOW-window2 includes four words and a BOW-window5 includes ten words for each target word. We compute the cosine similarity between the BOWs for the two target words to measure the similarity between them. Table 4 compares the results of the different models and shows that our approach outperforms the baseline results. However, our best performance, 0.52, is still lower than the best performance reported in previous work [Huang *et al.*, 2012], 0.66. One reason for this is that annotators did not distinguish semantic similarity and conceptual similarity, whereas our approach measures conceptual similarities with semantic similarity only acting as a constraint in mapping the words to concepts. For example, in the annotated dataset, *seafood* and *sea* have a high similarity score. CDC gives a low similarity score in this case because they are conceptually quite different. Nonetheless, the results show that considering the context of words in conceptualization improves word similarity judgements over naive conceptualization without context.

4 Sentence-Level CDC

For a short text with two or more instance words, some tasks require sentence- or phrase-level conceptualization. For example, in recommending an online advertisement based on a Web query, we need to compare the concepts in the advertisement with the concepts in the query. In this section, we describe our method to output a sentence-level concept based on word-level concepts. We propose two methods with different approaches for combining the individual word-level concepts. Two different experiments with a Bing search log, an ad-query similarity, and a URL title-query similarity, show the benefit of context-dependent sentence conceptualization.

4.1 Combining Concept Vectors

We describe our method to conceptualize sentences based on word-level concepts. In the previous section, we described our approach for conceptualizing each word-level instance with its context inferred by a topic model. To produce the sentence-level concepts, we directly extend the word-level conceptualization by combining the concept vectors of instances and propose the following two variations of doing so:

1. Combining with an equal weight (CDC-EQ) : We assume that each instance in the sentence has an equal contribution on the overall concept of the sentence. This assumption usually makes sense when the sentence is composed of few important keywords, such as ad keywords provided by the advertisers. We can compute a concept probability vector of an instance by Equation 2 and combine the resulting concept vectors with an equal weight as follows:

$$p(c|\vec{s}) \propto \frac{1}{|\vec{s}|} \sum_i p(c|s_i, z), \tag{4}$$

   where $\vec{s}$ is the instance list vector of sentence s and s_i is the word index of instance i. We can compute $p(c|s_i, z)$ with Equation 2.

2. Combining with inverse document frequency weights (CDC-IDF) : We assume that each instance in the sentence has its own importance on the overall concept of the sentence. With the intuition from information retrieval that words with high document frequencies are not as important as words with low document frequencies [Manning *et al.*, 2008], we multiply the concept vector of each word by IDF of the word to compute the overall concept of the sentence:

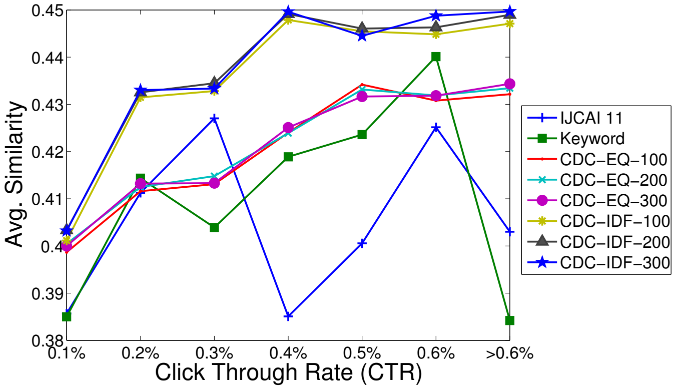
$$p(c|\vec{s}) \propto \sum_i \frac{1}{\text{DF}(s_i)} p(c|s_i, z). \tag{5}$$

The two variations are quite different in that CDC-EQ gives an equal weight on each instance, so an instance term with high document frequency contributes equally to the overall concept vector, but CDC-IDF prevents the general high document frequency words from contributing to the overall concept vector.

4.2 Experiment 3 : Advertisement Dataset

An important challenge in online advertising is in matching the ads to the users' information needs. An evidence of the ad-query match quality is whether users click on the sponsored URL in a search result page, so we look at the click through rate (CTR), a number of clicks on the advertisement divided by the total number of times the advertisement served. If we assume that users are more likely to click on the sponsored URLs for which their information needs match the advertisement, then one way to improve the CTR is to better match the ads to the queries.

We approach this ad-query matching task with sentence-level context-dependent conceptualization. For each sponsored-URL, there is a set of advertisement keywords from the sponsor, and we consider that as a sentence. We



Model	Correlation
Keyword	0.259
IJCAI 11	0.243
CDC-EQ-100	0.932
CDC-EQ-200	0.952
CDC-EQ-300	**0.955**
CDC-IDF-100	0.818
CDC-IDF-200	0.827
CDC-IDF-300	0.838

Figure 1: Correlations between binned click through rate (CTR) and average bid-query similarity. The graph illustrates the general pattern between CTR and average similarity. Our two proposed methods, CDC-EQ and CDC-IDF, reveal a much higher correlation between the two metrics compared with other baselines. The table on the right shows the Pearson correlation between the binned CTRs with the corresponding average similarities.

conceptualize these ad-bid keywords to derive the sentence-level concept vector and do the same for the users' search queries that resulted in the ads being served in the search results. Then we compute the similarity between the ad-bid concept vector and the query concept vector. If the correlation between this ad-query similarity and the CTR is high, then we can attribute it to successful sentence-level conceptualization.

For the experiment, we collect Bing search log of sponsored-URLs and queries with corresponding CTRs. We randomly select 68,016 sponsored-URL-query pairs from Aug. 01, 2012 to Aug. 31, 2012. We split these pairs into 7-bins with equal widths of 0.1% according to its CTR, and we aggregate the pairs for which CTR is more than 0.6% into one bin, and about 12% of pairs fall into this aggregated bin. The number of query-bid keyword pairs for which CTR is less than 0.6% is 59,680 (88%). With this dataset, we conceptualize the queries and the ad keywords and compute the cosine similarity between them.

Figure 1 shows the correlation between the average similarity of the ad-query pairs and the CTRs from our sentence-level CDC along with two different baselines. The *Keyword* method measures binary similarity in which if all of the bid keywords are included in the query, then the similarity is one, otherwise zero. The *Keyword* method shows high correlation with CTR except the aggregated bin. This result indicates the keyword matching method is acceptable for lower CTRs, but the method does not explain the relatively high CTRs for queries in the aggregated bin. CDC captures the relatively high similarity for the aggregated bin better than the keyword matching. Another baseline is a previous conceptualization method also based on Probase (*IJCAI11*) [Song *et al.*, 2011] which does not show a good correlation with CTR.

It is worth noting that CDC-EQ shows better correlations than CDC-IDF. The similarities with CDC-IDF are higher than CDC-EQ on average, but the Pearson correlation between the CTR and the average similarity of CDC-EQ method performs best. From this result, we can see that weighting each instance equally increases the correlation, meaning that even the ad-bid keywords with high document frequencies are important, perhaps because sponsors try to find the best words to describe the product.

4.3 Experiment 4 : Query Similarity

We describe an application of our sentence-level context-dependent conceptualization to computing the similarity between search queries and titles of the URLs clicked from the Bing search log data. Using click-through logs as implicit feedback improves the search performance [Joachims, 2002], so if we can better predict the Web pages that users will click, we can further improve search performance.

In a similar approach to the ad-query matching problem, we compute the concepts of URL titles and concepts of queries and measure the cosine similarity of the query-URL pairs. Just aiming for a high similarity score for a query and the clicked URL is not enough because some methods tend to generate high similarity scores for most query-URL pairs, so we instead aim for a significant difference between the similarity score of the query-URL pairs in the click-through logs and the similarity score of random query-URL pairs. Hence, we create a set of randomly sampled queries and URL titles and use it as a comparison set.

Table 5 shows the similarity results with various methods and different numbers of topics. The numbers in parentheses represent the similarity differences between the original query log dataset and the random comparison set. For the baseline methods, we use a previous conceptualization method (IJCAI11) [Song *et al.*, 2011] and the topic similarity between the query and the URL title based on LDA. The LDA-based approach generates the highest similarity scores for the query log, but it also generates high similarity scores for random pairs. This is because LDA reduces dimension of the queries and URL titles from the size of the vocabulary to a smaller dimension of topics for measuring the semantic relatedness. CDC-IDF achieves the largest difference between the similarities of the two datasets. Unlike the ad-query

Table 5: URL-Query similarity of click through data and random pairs. Numbers in the parenthesis indicate the differences between similarities of click through log and random pairs. CDC-IDF model shows the best differences. T# indicates the number of topics used for experiments

(a) Click through log

	IJCAI11	LDA	CDC-EQ	CDC-IDF
T100	0.31 (0.29)	0.55 (0.31)	0.39 (0.33)	0.42 (0.39)
T200		0.52 (0.31)	0.40 (0.34)	0.42 (0.39)
T300		0.50 (0.31)	0.40 (0.34)	**0.42 (0.39)**

(b) Random pairs

	IJCAI11	LDA	CDC-EQ	CDC-IDF
T100	0.02	0.24	0.06	0.03
T200		0.21	0.06	0.03
T300		0.19	0.06	0.03

matching, in this experiment, using different weights for the instances performs better. This is because many queries contain relatively meaningless terms such as 'search' and 'find'.

5 Conclusion

We described a framework for context-dependent conceptualization which combines the advantages of a Web-scale probabilistic knowledge base and a probabilistic topic model. The main result of our framework is the improved mapping of words in a sentence to context-sensitive concepts. We conducted experiments in which the context-sensitive approach to conceptualization yields better results for various tasks that require semantic understanding. Using word-level context-dependent conceptualization, we can predict unseen instances for frame elements which are valuable for semantic role labeling but require expensive human annotation. Context-dependent conceptualization can also identify word similarity in context. Sentence-level conceptualization using different weighting schemes can be used to match Web search queries and advertisements, as well as queries and URL titles. For both of these tasks, we used large real-world click-through logs to quantitatively evaluate our framework against baseline approaches. Through these experiments, we showed that our framework for context-dependent conceptualization is effective in various tasks that require deep understanding of short texts. Conceptualization is an important and general problem, and we showed a simple but effective framework to combine Probase and LDA. With recent advances in Web-scale, corpus-based probabilistic knowledge bases and probabilistic topic modeling, there is a great potential to make improvements based on our framework.

Acknowledgments

We thank Zhongyuan Wang and Wen Hua for constructive discussions and the anonymous reviewers for their helpful comments. Alice Oh is supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0006856).

References

[Baker *et al.*, 1998] C.F. Baker, C.J. Fillmore, and J.B. Lowe. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998.

[Blei *et al.*, 2003] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[Boyd-Graber *et al.*, 2007] J. Boyd-Graber, D. Blei, and X. Zhu. A topic model for word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1024–1033, 2007.

[Cambria *et al.*, 2013] E. Cambria, Y. Song, H. Wang, and N. Howard. Semantic multi-dimensional scaling for open-domain sentiment analysis. *IEEE Intelligent Systems*, 2013.

[Chemudugunta *et al.*, 2008] C. Chemudugunta, A. Holloway, P. Smyth, and M. Steyvers. Modeling documents by combining semantic concepts with unsupervised statistical learning. *The Semantic Web-ISWC 2008*, pages 229–244, 2008.

[Finkelstein *et al.*, 2001] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*, pages 406–414. ACM, 2001.

[Fürstenau and Lapata, 2009] H. Fürstenau and M. Lapata. Semi-supervised semantic role labeling. In *Proc. of EACL*, 2009.

[Gildea and Jurafsky, 2002] D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.

[Griffiths and Steyvers, 2004] T.L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5228–5235, 2004.

[Hua *et al.*, 2013] Wen Hua, Yangqiu Song, Haixun Wang, and Xiaofang Zhou. Wen hua, yangqiu song, haixun wang, xiaofang zhou. In *WSDM*, 2013.

[Huang *et al.*, 2012] Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the Association for Computational Linguistics 2012 Conference (ACL '12)*, Jeju, Republic of Korea, July 2012.

[Joachims, 2002] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 133–142. ACM, 2002.

[Manning *et al.*, 2008] C.D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

[Song *et al.*, 2011] Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen. Short text conceptualization using a probabilistic knowledgebase. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence*, pages 2330–2336. AAAI Press, 2011.

[Wang *et al.*, 2012] Jingjing Wang, Haixun Wang, Zhongyuan Wang, and Kenny Zhu. Understanding tables on the web. In *International Conference on Conceptual Modeling*, 2012.

[Wu *et al.*, 2012] W. Wu, H. Li, H. Wang, and K.Q. Zhu. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 International Conference on Management of Data*, pages 481–492. ACM, 2012.

[Yao *et al.*, 2009] Limin Yao, David Mimno, and Andrew McCallum. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 937–946. ACM, 2009.