# Named Entity Recognition in Query

Jiafeng Guo, Gu Xu, Xueqi Cheng, Hang Li - SIGIR 2009

A Summary by Alberto Ueda
CS Dept/UFMG - 2017/09

## 1 CONTEXT

Given a text query, such as "Harry Potter walkthrough", how to detect the named entities in the query ("Harry Potter") and classify them into different classes ("Game", "Tutorial")? To answer that, Guo et al. define a novel task called NERQ (for Named Entity Recognition in Query). They solve this task proposing both a probabilistic-based framework and an instantiation of it using a variation of LDA (Latent Dirichlet Allocation). This variation is called Weakly Supervised LDA (WS-LDA) and it is another contribution of the paper for the research community in query understanding.

Although the Named Entity Recognition (NER) task already exists, the authors claim that usual techniques for NER do not succeed in NERQ. The main reasons for this incompatibility are the short length of the queries (2–3 words on average, in comparison to large documents analyzed in NER) and the bad formatting of the queries (e.g., typos), while methods of NER usually rely on several features of documents written in natural language as input to machine learning approaches.

The work is inspired by the ideas presented by Paşca (CIKM, 2007), a named entity recognition method focused on offline query log mining. Paşca proposed a deterministic approach (referred as *Determ*) where each named entity is only associated to one class. In contrast, using a probabilistic approach, the work of Guo et al. allows the association of multiple classes to the same entity.

## 2 SETUP

- Real data set provided by a commercial web search engine. As two of the paper's authors work at Microsoft Research Asia, that search engine is probably Microsoft Bing™.
- Total of 6B queries, on which 930M were unique. 140K queries were processed and 400 of them were used in the evaluation.
- 4 classes (topics) considered: *Movie, Game, Book,* and *Music.*
- Sample of 400 pairs $[query, class]$ manually labeled as "correct" or "incorrect".
- Movie (111 queries), Game (108), Book (82), Music (99).
- 180 named entities from the websites of Amazon, GameSpot, and Lyrics. Four human annotators labeled the classes of the named entities.
- 120 *seed* named entities in training set and 60 named entities in test set. At the end of the algorithm, a total of 1.5M named entities were indexed.
- 70% of the sample queries contain at least one named entity. 1% of the sample has two or more named entities.

## 3 APPROACH

The task of recognizing and classifying a named entity of a given query is modeled as follows. Let $e$ be the named entity in the query, $t$ a context (other words of the query), and $c$ a class the named entity $e$ belongs to. In this way, any query can be represented by a group of triples $(e, t, c)$. Remember that the goal of NERQ is to detect the named entity $e$ and associate it the most likely class $c$ to $e$. Thus, this task can modeled as finding the triple $(e, t, c)$ with the highest probability for a given query. To compute these probabilities:

$$Pr(e,t,c) = Pr(e)\ Pr(c|e)\ Pr(t|e,c) = Pr(e)\ Pr(c|e)\ Pr(t|c)$$

In other words, we can compute how likely a given query is generated from triple $(e, t, c)$ if we know *a priori* the popularity of the named entity $e$, the likelihood of a class $c$ given $e$ and the likelihood of a context $t$ given $c$.

Finally, to estimate $Pr(c|e)$ and $Pr(t|c)$, there is need to learn a topic model, what can be done using methods such as LDA. However, to correctly align the latent classes of LDA and the predefined classes used in the manual labeling, the authors introduce a variation of LDA, the WS-LDA. They use the term "weak supervision" in the sense that the objective function used in learning is based on binary human judments ("correct" or "incorrect"), instead of probabilities in the range $[0, 1]$.

## 4 SIGNIFICANCE

The authors address an arguably real world problem within query understanding and use real data to evaluate their results. The probabilistic approach proposed, using query logs and a topic model, outperforms conventional methods such as Determ and classic LDA in terms of accuracy, overall class likelihood, and convergence speed on the training phase. Both the formal definition of NERQ and the probabilistic approach to perform the task can be useful to future improvements in query-based tasks (e.g., intent identification in web search, query suggestions). However, as the authors claim, this is only the first work focusing on this specific problem. The baselines considered are either focused on other task (Determ) or adapted to the NERQ task (LDA, actually an unsupervised method). Therefore, we need further studies on this task to correctly address the effectiveness of this approach.

## 5 NOVELTY

The authors define a novel task, NERQ, since standard NER techniques do not perform well when the text documents are queries. This definition can help future researchers on correctly addressing its efforts and choosing precise baselines. Likewise, the authors present the first approach to deal specifically with NERQ, combining a probabilistic framework with a topic model obtained by human-supervised LDA.

## 6 MAIN CONTRIBUTIONS

(1) A formalization of a novel task: NERQ.
(2) A probabilistic approach to perform the NERQ task.
(3) A supervised method for learning topic models: WS-LDA.