IDIAP RESEARCH REPORT

EXPLICIT SUGGESTION OF QUERY TERMS FOR NEWS SEARCH USING TOPIC MODELS AND WORD EMBEDDINGS

Parvaz Mahdabi Andrei Popescu-Belis

Idiap-RR-21-2016

AUGUST 2016

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Explicit Suggestion of Query Terms for News Search using Topic Models and Word Embeddings

Parvaz Mahdabi and Andrei Popescu-Belis

Idiap Research Institute
Centre du Parc
Rue Marconi 19
CH-1920 Martigny, Switzerland

August 2016

Abstract

This report presents a study on assisting users in building queries to perform real-time searches in a news and social media monitoring system. The system accepts complex queries, and we assist the user by suggesting related keywords or entities. We do this by leveraging two different word representations: (1) probabilistic topic models, and (2) unsupervised word embeddings. We compare the vector representations obtained by these two approaches to find related keywords (i.e. suggestions) with respect to specific queries, taken from the query log of a commercial system. Through crowdsourcing we solicited relevance judgments and compared the two methods. Our results show that word embeddings outperform topic models for keyword suggestion.

1 Introduction

With the recent upsurge in data, finding the relevant information with respect to a user query and filtering out the noise has become more and more difficult. The data deluge has transformed many businesses in the recent years, and there is an increasing demand for systems that help users aggregate, extract, and analyze information from different sources. In this work, we study one remedy to data deluge problem, which is assisting users with creating high quality queries, which have the potential to help users find more relevant information. We guide users in formulating their queries by suggesting additional words to include in a query, given a set of initial terms. The main goal is to add terms that disambiguate the query and possibly extend it towards the most relevant aspects of the information looked for by users.

The contributions of this paper can be summarized as follows:

- We design a crowdsourcing experiment to evaluate the keywords suggested to expand user queries. We work with a pre-selected set of query terms from the query log of a news and social media monitoring system, shown in Table 1 on page 5.
- We compare and evaluate two different word representations for the purpose of keyword suggestion, on four dimensions: relevance, specificity, relatedness, and predicted impact on search. The two word representations are word embeddings from word2vec and topic models from LDA.

2 Methods

2.1 Topic Modeling

Our objective is firstly to analyze and understand what topics are prominent in each document. Secondly, we use this analysis to improve the keyword suggestion component of our news recommender system, by determining: (1) what are the main topics related to each query, (2) which topics should receive more emphasis to obtain a better query representation, and (3) which terms from which topics can play important role in improving the query representation.

Probabilistic topic models are a well-known tool for unsupervised analysis of text, providing a latent topic representation of the corpus. We used the Online Variational Bayes for Latent Dirichlet Allocation [1], which proposes an approximate inference algorithm for calculating the posterior which is scalable for analyzing massive collections and has constant memory requirements thanks to a stochastic optimization with a natural gradient step.

After performing the inference task in LDA and calculating the topic mixture proportions and the topic assignments for each word, these approximate probability distributions are used for calculating candidate terms for keyword suggestion.

To compute a *term relevance score*, we proceed as follows. For a query Q composed of query terms $\{q_1, q_2, ..., q_i\}$, the objective is to find terms from the related topics to the query. Using LDA, we determine the distribution over the topic $z = j$ of each word w, noted $p(w|z = j)$ from the training data. The procedure is described in Equation 1. We used the independence assumption between query terms and the chain rule to obtain the final quantities in this equation. Bayes' rule is used to obtain $p(z = j|W)$ from the calculated topic-word probability distribution.

$$p(w|q_1...q_i) = \sum_{j=1}^{T} p(w|z=j)p(z=j|q_1...q_i) \tag{1}$$
$$= \sum_{j=1}^{T} p(z=j|w)p(z=j|q_1)...(z=j|q_i)p(z=j)^{i+1}$$

The term relevance score is calculated for all terms in the vocabulary. The suggested terms are ranked according to the term relevance score, and the top k terms are suggested as relevant concepts to the user. The size of the vocabulary is denoted by the product of number of topics and the number of terms selected from each topic. In the implementation, we bring these values to log scale and rank terms according to the log of the term relevance score.

We build the topic model from a dump of Wikipedia articles in English (dated August 2015). The number of topics is set to 2000 for the model trained on Wikipedia articles. The number of top words selected from each topic is set to 100. The hyper parameters affecting the sparsity of the model, alpha and beta, the priors are set to 1/number of topics. The number of iterations is set to 1000.

2.2 Word Embeddings

Neural word embeddings can learn meaningful representations of words in geometrical dimensions. A good embedding provides vector representations of words such that each word vector captures syntactic and semantic features of the word, while the relationship between two word vectors is related to some aspects of the linguistic relationship between the two words. Cosine similarity has been used as a common metric to calculate distance between words in an underlying embedding space where meaning is interpreted through geometry [3]. Furthermore, word vectors can be optimized to obtain any desired property with respect to a task, such as named entity recognition, semantic role labeling, or machine translation [2, 4].

In this work, we use the *word2vec* package. There are two word vector methods in this package. (1) The first method averages the hidden representation of a context to predict the middle word (CBOW model). (2) The second method averages the hidden representation of the middle word to predict the surrounding context (Skipgram model) [3]. In this work we use the Skipgram model.

We extract bigrams and trigrams using collocation analysis. These bigrams and trigrams are calculated as follows:

$$\text{score}(w_a, w_b) = \frac{\text{count}(w_a w_b) - \delta}{\text{count}(w_a) - \text{count}(w_b)} \tag{2}$$

Bigram $w_a w_b$ gets added to the vocabulary on condition that its calculated score surpasses a threshold of 10. This process is performed recursively to obtain trigrams. This is similar to what is used in [3].

We build a weighted average vector from the query terms, giving a weight of 0.5 to terms in OR clauses, 1.0 to terms in AND clauses, and -1.0 to terms in the NOT clause. An example query is shown in Section 3.2. We then calculate cosine similarity between the weighted avg. query vector and all other entries (i.e. terms) in the vocabulary. We retrieve the most 10 related terms for each query.

We used 300 dimensional feature vector, and used a window of 5. The number of negative samples is set to 10. We perform downsampling of very frequent terms in the process of learning word embeddings.We keep a vocabulary of 2 million entries. We used the Skipgram implementation of the gensim toolbox[1].

3 Experimental Setup

3.1 Dataset

We use the following dataset in our experimental setup: the English dump of Wikipedia obtained on August 2015. This is used to train both the topic model and the word2vec approaches to find word vector representations. The dataset is composed of nearly 5 million documents.

3.2 Query Representation

The system allows users to construct Boolean queries composed of keywords and phrases, which are represented in Conjunctive Normal Form (CNF). A CNF formula is an "AND of ORs" (i.e. a Boolean conjunction of clauses, where a clause is a disjunction of literals). While CNF queries are mainly used by expert searchers such as librarians, lawyers, or patent searchers, the system considered here assists non-specialist users with query construction through a user-friendly interface which provides an intuitive view of AND/OR connectives and can recommend related terms (concept names) for each field. An example query in CNF form is as follows:

> (FedEx OR Parcel Post OR Postage Stamp OR Royal Mail OR United Parcel Services OR United States Postal Service OR Universal Postal Union) AND (Privatization OR Private Sector OR Public Sector OR Postal Services)

[1] https://radimrehurek.com/gensim/

Table 1: Queries used in our evaluation.

Id	Query
1	OPEC
2	(FedEx OR Parcel post OR Postage stamp OR Royal Mail OR United Parcel Service OR United States Postal
	OR Universal postal union) AND (Privatization OR Private sector OR Public sector OR postal services)
3	(Mail OR Post OR Postal OR Postal industry) AND (E-commerce OR Electronic commerce OR Online shopping)
4	(Digital Video Broadcasting OR Digital television OR Digital video) AND Over-the-top content
5	(Micorfinance OR Remittance OR mobile money) AND Financial services
6	(Cartography OR GIS OR Geocoding OR Geographic information system OR Spatial analysis OR Visualization)
7	Android AND mobile AND operating system
8	Technological singularity
9	(E-commerce) AND (Innovation OR inclusion OR integration)
10	(Analytics OR Recommendation engine OR Visualization OR classification OR clustering)
11	(parcels) AND (Customs OR Trade facilitation OR postal)
12	(mobile device OR mobile phone OR smartphone) AND (mobile game OR social network game OR video game)
13	(cell nucleus) AND (3d cell imaging OR bright field microscopy OR microscopy)
14	(privacy) AND (big data OR internet of things) AND (algorithm OR machine learning OR real-time computing OR
	data visualization OR data analysis OR analytics OR visualization)
15	Diamond industry

Table 1 presents several queries obtained from the query log of the system. These queries will be used to evaluate the performance of the proposed methods in the rest of this paper.

3.3 Using Human Judgments of Suggested Keywords

In this section, we propose a task that creates a formal setting where humans can evaluate four aspects of the quality of the suggested keywords.

In the beginning of the task, a human subject is presented with the Boolean query (composed of one or more query terms) alongside a textual representation of the query. The task then presents to the subject the assisted search scenario and the overall goal of improving the query formulation using suggested keywords.

The task starts with a control question to verify whether the subject is

familiar with the suggested word or phrase. The control question specifically tests whether the subject is able to recognize the correct definition of the suggested keyword among multiple proposed definitions.

For each word or phrase presented to the subject for a given initial query, the subject is asked to estimate four aspects of quality. First, the subject evaluates whether the association between the suggested keyword and the query terms is relevant. Then the subject identifies the level of specificity or generality of the suggested term compared to the query terms in the initial query. The subject then estimates to what extent the suggested keyword is related to the query term. Finally, the subject is asked to guess whether adding the suggested term to the query can improve the results expected from the query.

In our study, the relevance of the suggested keywords to the query is assessed using a 3-point scale: {not-relevant, somewhat relevant, very relevant}. The specificity of the suggested keyword with respect to the other words of the query is evaluated using four levels: {more specific, synonym, same level, more general}. The relatedness of the suggested term to the query is evaluated with a binary rating. Finally, the expected impact of the suggested keyword on the search results of the expanded query is also evaluated with binary levels.

3.4 Performing the Human Evaluation

The task described above was offered on the Amazon Mechanical Turk platform, which allows workers with no special training or knowledge to perform small jobs for a small reward. We recruited four workers per evaluation task, i.e. evaluating a suggested keyword with regards to a query. Each worker spent on average 30 seconds answering the task. Workers were paid 0.10$ per task and we selected workers that obtained an approval rate of higher than 99% according to their performance on previous tasks.

We calculated inter annotator agreement for 4 raters and 150 keywords (or keyphrases), and compared the topic model with word2vec. The values of Fleiss' Kappa, Z, and p-values are shown in the following four tables. This agreement is quite moderate, and this is why we used the average rating for each keyword.

Table 2: Inter annotator agreement according to Fleiss Kappa for relevance dimension, 4 raters, 150 suggested keywords for 15 queries (10 per query)

Method	F. Kappa	Z	p-value
W2V	0.248	10.4	0
TM	0.241	9.68	0

Table 3: Inter annotator agreement according to Fleiss Kappa for specificity dimension, 4 raters, 150 suggested keywords for 15 queries (10 per query)

Method	F. Kappa	Z	p-value
W2V	0.143	6.36	2.02e-10
TM	0.108	4.51	6.38e-06

Table 4: Inter annotator agreement according to Fleiss Kappa for relatedness dimension, 4 raters, 150 suggested keywords for 15 queries (10 per query)

Method	F. Kappa	Z	p-value
W2V	0.262	7.86	3.77e-15
TM	0.165	4.92	8.57e-07

Table 5: Inter annotator agreement according to Fleiss Kappa for impact on search dimension, 4 raters, 150 suggested keywords for 15 queries (10 per query)

Method	F. Kappa	Z	p-value
W2V	0.202	6.07	1.31e-09
TM	0.199	5.94	2.8e-09

4 Experimental Results

We prepared the data for human subjects to review from the two models described in Section 2 for suggesting keywords to expand a query. Both models are trained on English articles of Wikipedia. The results of the comparison of the two proposed methods (W2V and TM) are presented in Figure 1.

We are primarily interested in comparing W2V and TM in terms of relevance of the suggested keywords, as shown in plot (a). According to our crowdsourced results, W2V is able to suggest a higher percentage of somewhat relevant and very relevant keywords compared to TM. In plot (b) of Figure 1 the analysis over the specificity dimension is presented. According to the obtained judgments, W2V continue to suggest more specific keywords, while the TM suggestion are more general. In plot (c) of Figure 1, we can see that the judges evaluated the suggested keywords of W2V achieves higher percentage of relatedness with respect to suggestions of the TM method. Finally, in plot (d) of Figure 1, the judgments indicate that W2V is able to suggest keywords that have a better impact on improving the search results.

In total, we can see that W2V achieves a higher performance compared

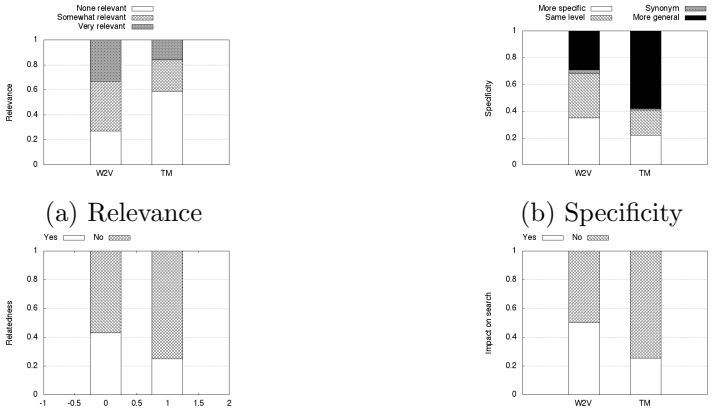
to TM, by suggesting more relevant and more specific keywords. The terms suggested by W2V are evaluated by judges as being more useful as expansion terms, i.e. potentially leading to an improvement of search results. The judgments from the crowdsourcing experiments are also presented in Table 6.

Table 6: Results of the evaluation on 15 queries.

Evaluated aspect	**W2V**	**TM**
Relevance	0.73	0.41
Specificity	0.71	0.42
Relatedness	0.44	0.25
Impact on search	0.50	0.25

Several reasons may explain why the performance of Topic Models is lower than that of word2vec:

- Phrase extraction is done differently in the two models. In TM we extract all the titles of Wikipedia pages and add them to our vocabulary, while in W2V we use collocation analysis to detect important bigrams and trigrams and then we enhance the vocabulary with these detected phrases.
- In TM we consider a smaller vocabulary compared to W2V for tractability. The vocabulary of TM is includes the 100 first terms of each topic.



(a) Relevance

(b) Specificity

(c) Relatedness

(d) Search impact

Figure 1: Results of the crowdsourcing study.

5 Conclusion

In this study, we showed that query formulation can be effectively assisted by suggesting relevant keywords to the user using recent advances in distributional semantics. We experimented with probabilistic topic models and word embeddings (word2vec skip-gram with negative sampling). The best results in terms of relevance, specificity, and impact on search were achieved by the skip gram model. Both our models were trained using the English Wikipedia articles.

Future work should consider the usage of the suggested keywords in the context of query expansion. It would be important to compare the two approaches by looking at the retrieval accuracy actually achieved when using the suggested keywords for query expansion. It is also possible to cluster the suggested keywords into meaningful groups, for an improved user experience.

Acknowledgments

This work was funded by the Swiss Commission for Technology and Innovation CTI/KTI.

References

[1] M. Hoffman, D. Blei, and F. Bach. Online learning for latent dirichlet allocation. *NIPS*, 2010.

[2] T. Luong, R. Socher, and C. D. Manning:. Better word representations with recursive neural networks for morphology. In *CoNLL*, pages 104–113, 2013.

[3] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *In Proceedings of NIPS*, 2013.

[4] W. Y. Zou, R. Socher, D. M. Cer, and C. D. Manning:. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*, pages 1393–1398, 2013.

6 Appendix: Examples of Suggested Terms

The suggested terms presented in the two tables below correspond to the queries presented in Table 1, identified by their number between 1 and 15.

Table 7: Suggested terms obtained from W2V on 15 queries.

Q.	Terms suggested by word embeddings (W2V)
1	petroleum exporting, per barrel, oil exports, oil prices, mmb, price increases, embargo against, foreign debt, arms embargo, crude oil
2	regulatory authority, airwork, farm credit, postal system, banking services, cupe, marine insurance, private contractors, investment funds, militia pay
3	reseller, internet banking, services provider, call centres, online gambling, banking services, customer care, helpdesk, direct mail, postal services
4	sdtv, high definition, hdtv, multi channel, audio channels, isdb, multichannel, digital audio, digital satellite, stereo sound
5	mortgage loans, savings account, underwrite, banking services, internet banking, retail banking, institutional investors, investment funds, pension funds, investment opportunities
6	mathematical modeling, geospatial, spatial data, statistical techniques, datasets, statistical data, graphing, optimisation, summarization, digitization
7	device drivers, hypervisor, bootable, symbian os, windows operating, boot loader, qnx, userland, thin client, ibm compatible
8	teleology, magical thinking, cognitive dissonance, scientific progress, memetics, cosmic inflation, biological evolution, paradigm shift, reductionist, evolutionary process
9	technological innovations, technological development, development, emerging technologies, environmental sustainability, widespread adoption, dsdm "Dynamic System Development Method", cross disciplinary, continuous improvement, marketing strategies
10	optimizer, skills assessment, summarization, arcgis, image compression, rapid prototyping, code data, bootstrapping, rdbms, toolset
11	business transactions, excise act, overseas trade, investment opportunities, laws regulating, economic sectors, imported goods, bilateral trade, regulatory agencies, global markets
12	multiplayer games, online multiplayer, wireless devices, nintendo wi, multiple platforms, portable devices, cellphones, handhelds, multiplatform, sony psp
13	microarrays, electron microscopy, electrophoresis, biopolymers, photoelectric, cytogenetics, photosensitive, ray diffraction, gel electrophoresis, thin films
14	mathematical modeling, statistical techniques, external resources, spatial data, computing hardware, summarization, computer vision, relational databases, parallel computing, optimisation
15	manufacturing industry, cottage industry, textile manufacturing, diamond mining, semiconductor industry, vertically integrated, petrochemicals, manufacturing, oil gas, handloom

Table 8: Suggested terms obtained from TM on 15 queries.

Q.	Terms suggested by topic model (TM)
1	pipeline, field, planted, electric, gas, fuels, company, petroleum, energy, oils
2	address, markings, e-mails, postal, services, message, spam, email, posted, mail
3	appointed, newspaper, primes, london, generations, member, canadian, magazine, president, minister
4	library, systems, cinema sound system, television, information, video, audio, camera, cable, internet
5	cables, access, phone, orleans, gundam, telephones, wireless, telecommunication, internet, services
6	stories, radar, filmed, acacia, targets, fairies, version, books, brother, manning
7	services, agencies, afghanistan, internet, director, member, generations, staff, wireless, commissioner
8	earls, lords, systems, labs, universities, center, baron, field, scientist, scientific
9	england, english, kings, bishop, earls, euros, european commission, treaties, access, services
10	new york, cities, brooklyn, manhattan, library, cylinder, species, coach, rear, rpm
11	message, mail, e-mails, posted, workers, vice, browns, gta, labor, unions
12	players, songs, single, band, musical, album, sounds, levels, mode, fighter
13	stadium, arena, blue, armed, homes, football, redding, teams, yard, sports
14	max, information, media, manning, webbing, philosophers, ethics, world, lovecraft, benson
15	album, feat, rappers, produced, factories, steele, gems, works, manufacture, planted