International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

Some Issues in Application of NLP to Intelligent Information Retrieval System and Guidelines for its Solution

Shrey Patel
B.E. Computer Engineering, Gujarat Technological University, Ahmedabad, Gujarat, India

ABSTRACT

The process of getting the semantic information out of vast text data available is not easy. Many of the recent work on Intelligent Information Retrieval (IIR) dealt with usage of natural language processing (NLP), but the results were not so encouraging. Even currently applicable state-of-the-art NLP gives only moderate results when used in document retrieval systems. Firstly, this research paper addresses the key issues that occur when incorporating NLP techniques in IIR. We look in detail, what are the causes of issues. Then we propose some solutions to tackle with these issues, for getting better IIR system outputs.

Keywords: Natural Language Processing, Information Retrieval, Document Retrieval, Word Sense Disambiguation, Query, Machine Learning

I. INTRODUCTION

An Information Retrieval system is one that searches a collection of documents, and retrieves exactly the set of documents that best satisfies a user's query. An example of this is searching the World Wide Web using a search engine like Google, Bing, etc. These search engines take the user's query (which is in the form of natural language), and ranks the documents accordingly. Then it returns the set of webpages that best matches with the query. The inputted query can be in the form of a sentence, or just keywords. IR systems work in 2 steps: 1) Indexing 2) Matching. Many models exist for indexing. Among these, Vector Space model is widely used and it works as a bases for many extended versions. A model's accuracy is largely determined by the weighting factor used.

Natural Language Processing, or NLP, deals with understanding and manipulation of unstructured text, which is in the form of human language. NLP is widely used in IR applications relating to Question-Answering, Information Extraction, Machine Translation and automated summarization. However, current application of NLP techniques in Document Retrieval (which contains large amount of text) is found to be moderate [1], [2]. Recent research focused on the usage of various NLP techniques like parsing, chunking, word sense disambiguation etc. to document retrieval had similar results as compared to simple statistical IR. NLP is mainly used in IR task relating to indexing documents. Other methods are used in document matching part of IR system. This research paper focuses on issues relating to the application of NLP in IR, and propose solutions that takes the advantages of NLP techniques for relatively better IR system outputs.

II. CURRENT ISSUES WITH NLP POWERED IR

This section shows the problems faced when using some of the key NLP techniques to IR application. It also suggests solutions to the mentioned issues.

A. Understanding of input query

It is very important to understand the semantics of input query, since the context of query can be largely determined from it. The Length of query plays a significant role for the use of NLP in query processing. A common intuition is that shorter query contains less contextual information as compared to longer queries. Below Figure 1 displays the data length for different IR systems.

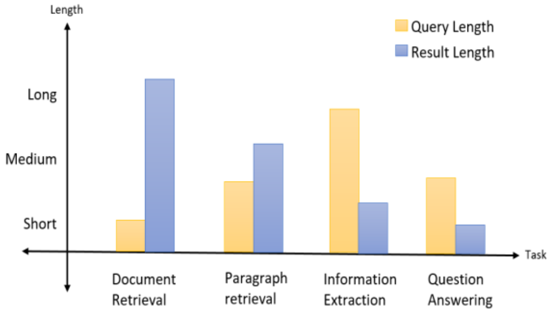

Figure 1

One more thing can be noted is the corresponding length of output results. In task like Question Answering (QA) system, the result is quiet short, typically just an answer in a few sentence. So, the information content in such QA results is dense, thus requiring more refined processing. On the other hand, task like document retrieval has larger results. Document retrieval system gives a collection of documents ranked according to the usefulness of results. For this, the system has higher probability that the requested information is contained in the top ranked documents. In the next section, we will see query expansion method that can be used to deal with short queries in document retrieval task.

B. Word Sense Disambiguation

Word Sense Disambiguation (WSD) is a task that identifies the meaning of a word in a sentence, when that word has different meaning in different context. For example, the word 'bank' refers to a financial organization in the sentence: "the loan was approved by the bank." Whereas, for the sentence: "willows lined the bank of the stream" refers to the land alongside a river. The use of WSD in IR systems like search engine is to improve the relevance of retrieved results (documents). An approach to use WSD in IR is by replacing the terms in the document vector by their senses (use in Vector Space Model). Vossen, P. 1998[3] finds that for the data relating to medical domain, average precision decreases by 23% for English, when using EuroWordnet (Wordnet is a popular sense inventory, that encodes various concepts and senses, and is widely used for disambiguation). However, when MeSH (Medical Subject Heading) was used on same data, they saw better improvements for English as well as German language. This can largely be attributed to the fact that domain specific ontology plays a major role for improving IR system performance (Also, see Thorsten Brants[1]).

So, for building an effective WSD component, we need a proper domain specific annotated corpora, or a specific thesauri pertaining to a domain. Unfortunately, such resources are not easily available, and does not even exist for many domains. Also, the real challenge is for general purpose (domain less) Information Retrieval Systems. Recent research on machine learning has gained a lot of attention for the task of WSD.

Following are the approaches used in WSD:

1. Dictionary / Knowledge Based Methods: These methods primarily use thesauri, or lexical knowledge bases. It does not use training corpora.

2. Supervised Learning: These methods make the use of sense-annotated training corpora, for initial training purpose. After this, it can further classily unseen cases. This method is not suitable for general – purpose IR systems, because such annotated corpora knowledge does not exist except for a small number of domains.

3. Semi-supervised Learning: These methods initially makes the use of small annotated corpora as initial basic training data, and then independently learns to classify senses from new data. The initial failure rate is high, but it gradually improves. This makes it unsuitable for general purpose IR systems. But it can be used to build domain specific IR systems, where moderate errors are acceptable.

4. Unsupervised Learning Method: This method assumes that "similar senses occur with similar contexts." It uses clustering method to classify words used in a context by using some measures of similarity between different contexts. This task is also referred to as word sense induction. These methods generally give less performance as compared to supervised methods, but do not suffer from knowledge-acquisition problems, as they are not dependent on training corpus.

From above mentioned 4 approaches, unsupervised learning seems to best fit with domain-less general purpose information retrieval like search engines. Current research in unsupervised machine learning is focused on improving performance of word sense induction systems.

C. Chunking

Chunking, or shallow parsing, deals with identification of part of speech and short phrases. Simple part-of-speech tagging tells us about the word category – it's relationship with adjacent words in a sentence; i.e. separating out nouns, verbs, adjectives etc. Sometimes, it is often required to get more information out of given query. One such application of chunking is named-entity recognition, which deals with extracting named-entity and it's type from query. For example,

"The {Indian Airways}$_{\text{company}}$ jet is set to take off to California at {14:15}$_{\text{time}}$" .

Here, Indian Airways and 14:15 are recognized as airway company and time respectively. This type of application is not necessary for a simple IR system such as document retrieval, but it is required for a different kind of task like QA-system or information extraction. Chunking helps to get a lot of semantic information. Also, chunking can be preferably used in place of n-grams during stemming, if it shows equal or more efficiency.

III. PROPOSED SOLUTIONS

As we have seen in the previous section, there are many issues arising because of the nature of task relating to information retrieval, and because of several constraints present in the NLP techniques. The approach of directly applying NLP techniques to IR systems, in a way that makes IR an *application of NLP* does not yield better results. We must take an integrated approach to build intelligent information retrieval system in order to get efficient retrieval. There are many measures available for the performance evaluation of a Retrieval System. Basic methods include calculating precision (P) and Recall (R).

Precision is the portion of retrieved documents that are relevant to the user's query:

$$\text{Precision} = \frac{|\{relevant\ document\}| \ \cap\ |\{retrived\ document\}|}{|\{retrived\ document\}|}$$

Recall is the portion of the total relevant documents that are retrieved by system.

$$\text{Recall} = \frac{|\{relevant\ document\}| \ \cap\ |\{retrived\ document\}|}{|\{relevent\ document\}|}$$

In order to evaluate the performance of document retrieval that ranks the retrieved documents based on relevance, Mean Average Precision (MAP) is used. Mean Average Precision is generally calculated for a set of queries. It is essentially the Mean of the *Average Precision Scores*, which is calculated on each query.

$$\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AP}(q)}{\text{Q}}$$

Where,

Q = number of queries

AP(q) is the Average Precision, defined as:

$$\text{AP} = \sum_{k=1}^{n} \text{P}(k)\Delta\text{r}(k)$$

Where,

k is rank in the sequence of retrieved documents

n is the number of retrieved documents
P(k) is the precision at cut-off k in the list
Δr(k) is the change in recall from items k-1 to k

1. Query Expansion

As we observe that many times, the length of query is quit short for document retrieval systems and we get fewer opportunities to explore contextual and semantic information from it. One approach to deal with this is by using query expansion. In this approach, this user's query is reformulated by including new keywords to the original query while indexing. These new keywords are formulated by finding synonyms, stems, checking for spelling errors etc. and then weighting the terms appropriately. Use of query expansion increases the total recall of the system, at an expense of decrease in precision. Many research suggest that the query expansion can potentially increase precision, if we include result set pages that are more relevant to user's query (by using some ranking function). In fact, many commercial search engines use ranking scheme like Tf-idf (Term frequency – inverse document frequency) along with query expansion to get better quality in search results despite of high recall. In any case, the problem of getting better results becomes dependent on ranking scheme used. An open research issue is when to use stemming (to improve results), and when to not use (to prevent ill-effects like low precision).

2. Semi-supervised Learning for sense disambiguation:

Semi-supervised learning technique makes the use of small size of sense annotated corpora for initial training phase. It then learns by itself from real data, and system improves over time. This property makes it very suitable for WSD in a domain specific IR, where comparatively little amount of training data is available. [4] Shestakov, D. presents an idea of *an intelligent web crawling* scheme that can adaptively gather corpus relating to a specific context/domain. Such a web crawler can be used to gather corpora of a specific domain of interest, make few annotation of senses or a related thesaurus, and then implement semi-supervised learning algorithm for mapping the senses throughout. This can lead to quick development of an IIR system focused on that domain. However, we are not aware of any such Intelligent Web crawling scheme that is applicable to this idea.

3. Use of Semantically Relatable Sets (SRS):

SRS of a sentence is a collection of unordered words of that sentence, that appear as linked nodes of the semantic graph. For example, SRS of the sentence:
"*The neighbour bought a new phone from store*"
Might look like this:

a. {The, neighbour}
b. {neighbour, bought}
c. {bought, phone}
d. {new, phone}
e. {bought, from, store}
f. {a, phone}

[5] Mohanty et. al. defines SRS and proposes method for automatically generating and linking it. Use of SRS based search technique gives very high precision, thus making it very attractive for it's use in document retrieval. It's few enhanced versions also overcome the issue of having low recall. Use of SRS drastically improve the result quality, and is also successful due to highly accurate methods available to automatically extract such sets.

IV. CONCLUSION

In this paper, we saw some important issues relating to the use of NLP techniques in Information retrieval process, and studied its causes in detail. These issues arise mainly because of the nature of the task of IR system relating to document retrieval, and because of several shortcomings of NLP techniques present till date. We framed some important guidelines and solutions to deal with these issues. At last, we saw some approaches that can be used to make the task of document retrieval more accurate in terms of relevance, and semantics.

V. REFERENCES

[1]. Brants, T. (2003, September). Natural Language Processing in Information Retrieval. In CLIN.

[2]. Voorhees, E. M. (1999). Natural language processing and information retrieval. In Information Extraction (pp. 32-48). Springer, Berlin, Heidelberg.

[3]. Vossen, P.(1998) EuroWordNet: a multilingual database with lexical semantic networks, Kluwer Academic Publishers, Norwell, MA, USA.

[4]. Shestakov, D. (2014). Intelligent Web Crawling. Intelligent Informatics, 5.

[5]. Mohanty, R., Dutta, A., & Bhattacharyya, P. (2005, May). Semantically relatable sets: building blocks for representing semantics. In MT Summit (Vol. 5).

[6]. Voorhees, E. M. (1993, July). Using WordNet to disambiguate word senses for text retrieval. In Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 171-180). ACM.

[7]. Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., & Zavitsanos, E. (2011). Ontology population and enrichment: State of the art. In Knowledge-driven multimedia information extraction and ontology evolution (pp. 134-166). Springer Berlin Heidelberg.

[8]. Petasis, G., Vichot, F., Wolinski, F., Paliouras, G., Karkaletsis, V., & Spyropoulos, C. D. (2001, July). Using machine learning to maintain rule-based named-entity recognition and classification systems. In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (pp. 426-433). Association for Computational Linguistics.