Proceedings of the Federated Conference on Computer Science and Information Systems pp. 45–48
DOI: 10.15439/2020F113
ISSN 2300-5963 ACSIS, Vol. 21

Concept Blueprints Serving More Focused User Queries

Kurt Englmeier
Schmalkalden University of
Applied Science, Blechhammer,
98574 Schmalkalden, Germany
Email: k.englmeier@hs-sm.de

***Abstract*—Information Retrieval is about user queries and strategies executed by machines to find the documents that best suit the user's information need. However, this need reduced to a couple of words gives the retrieval system (IRS) a lot room for interpretation. In order to zero in on the user's need many a IRS expands the user query by implicitly adding or explicitly recommending the users further useful terms that help to specify their information need.**

Queries often do not comprise more than a handful of terms, which, in turn, do not sufficiently represent the user's need. In this paper, we propose and demonstrate an approach that enables users to resort to implicitly more complex query expressions. We call these semantic structures concept blueprints. Furthermore, users have the possibility to define the blueprints on their own. The purpose of the blueprints is to spot more precisely the text passage that fits the user's information need.

I Introduction

INFORMATION Retrieval (IR) is the process of looking up documents that suit the information need of the user or, in other words, that are relevant to the query terms expressed by the user. The more detailed the search query, the better the retrieval results. Therefore, IRS usually encourage users to add further query terms from a list of recommended terms that may also address the context of their query. The recommended terms happen to appear together in texts in close proximity or have been selected together previously by other users presumably having the same information need.

By looking up documents whose content is best summarized by terms that match the query terms the IRS supposedly provides the user with the required information. The terms summarizing the document's content and the ones representing the query must be somehow similar.

A query "long-term consequences covid-19 infection" will quite likely lead us to the information we are looking for, because the query terms appear in one form or another (e.g. as synonyms) in the retrieved texts. We probably will be satisfied with the documents provided.

Things are slightly different with a user query "covid-19 infections Paris yesterday". We may get statistics about Covid-19 infections including detailed figures for Paris. If we are lucky, we find yesterday's figures for the French capital in one of the retrieved documents. However, many retrieval results may not mention this particular figure we are looking for. One may think, it's a bit strange to use the term "yesterday" in query. Our retrieval experiences tell us that this term may not be quite useful for a successful search.

In other situations, things are not so obvious. Querying Google about the "global average runtime of nuclear power plants" provides mainly statistical information that enables you to calculate the answer yourself. Your query results in useful data around the information you need, but it takes you a lot of time and effort to scan through all the documents provided and to produce the answer you require.

There is a useful document available (also on the web) answering exactly your question *in one of its paragraphs* (see figure 1). However, you won't find the corresponding document among the first thirty something retrieval results.

> As a result of the decline in new nuclear power plant construction, the global nuclear power fleet is becoming increas-ingly outdated. In July 2019, the average age of the world's reactor fleet was 30 years, in other words three-quarters of the approximately 40-year service life that plants are generally designed for. Assuming a service life of 40 years, by 2030 another 207 reactors will have been taken off the grid (those that went online between 1979 and 1990) and a

Fig. 1. Section of a text and its representation after basic text patterns have been identified and accordingly annotated.

The problem results from the typical design of information retrieval processes. In short, all documents of the data source are indexed using the weighted index terms according to their relevance for the content of the *entire* document. User queries are matched against these index terms, and the documents with highest relevance values rank top in the result list provided to the user. The relevance value depends on the content of the entire document. The relevance of a single chapter in a document is blurred by the overall relevance value and term list of the document.

So far, the problem is well-known and barely spectacular. Search engines just work this way. In principle, text

classification and text mining adopt the fundamental methods of information retrieval.

The work presented in this article reflects the current state-of-work of the research group of the Schmalkalden University of Applied Science. The prototype applies supervised learning for a semi-automatic approach to extract, distill, and standardize data from text. Even though the prototype shown here still represents work in progress, it demonstrates its potential in the detection of fake news and misinformation.

II. Related Work

Our approach is designed around the paradigm of fact retrieval emphasizing natural language [1, 2, 3, 4] and the support of users in constructing more complex search queries [5, 6]. It is based on a combination of Named Entity Recognition (NER), Bag of Words (BoW), and Word N-Grams [7, 8]. We assume that a specific combination of keywords and annotated numeric expressions uniquely reflects a particular fact.

We can imagine a variety of theme-specific BoWs (for locations, names, expressions of aggression etc.) applicable in our context together with Named Entities for common patterns in text reflecting time, amounts, distances, and the like. This process usually combines key words and common text patterns. Finally, each pattern is annotated by an appropriate term that summarizes the meaning of the pattern.

Generic named entities help to standardize factual information and to abstract away the different forms of expressions for essentially the same thing. However, it does not suffice just to annotate generic patterns. We can also easily imagine that Named Entities may relate to ontologies that serve specific interpretation or calculation purposes.

NER in the context described here operates with BoWs addressing locations, persons, organizations, or institutions (Wall Street, Dow Jones, White House, Bangladesh, for instance). Furthermore, we use key words such as “Mr.” or “Health Senator” that hint to names of persons. The system takes these names and feeds them into the respective bag of words.

There are further interesting key terms pointing to names. For example, the term “by” following the title of an article leads the list of names authoring that article. The identification of proper names benefits from the analysis of sequential dependencies when bags of words can be produced automatically instead of manually. There are promising approaches to automatically identify names (and other important key expressions) in texts using conditional random fields (CFR) [9] or hidden Markov Models (HMMs) [10]. Inclined to CFR, we integrated a feature that proposes, for example, all names starting with capital letters and followed by an abbreviation as organization names, such as National Institute of Health (NIH) or Korean Electric Power Corporation (KEPCO).

The identification of facts starts with information extraction [11] and the annotation of the extracted text pieces according to the meaning they express [12]. Annotation has two roles: first, it adds a meaningful term to the extracted text, in particular the numeric data. Such patterns, for example, represent dates, percentages, numerical data, distances, and the like. Second, the annotations (and keywords) from the first annotation are further annotated. This process (if iteratively performed) produces an increasingly more abstract representation of the text and numeric data in the text piece under consideration. Semantic markers [13] are the smallest fraction of a text covering a certain meaning discernable from the other fractions. Together they mark the meaning of a particular piece of text.

III. The Paradigm of Concept Blueprints

For a more fine-grained retrieval that spots only the most relevant text sections in all documents, the classic IR approach needs to be adjusted. Application areas of such a type of retrieval are finding and extracting particular facts from texts of a collection or locating text sections that are pertinent for a particular situation manifested in a balance sheet, service request, or claim. This form of information retrieval has a prominent place in Legal Technology (LT), for instance.

To serve such a request, we have to modify the classical information retrieval process and integrate additional functionalities adopted from fact retrieval:

- Retrieval on chapter or sentence level
- Extraction of relevant text sections
- Standardized and contextualized representation of facts
- Special consideration of numerical information
- Inclusion of basic inference mechanisms

The central element in our approach for a combination of text and fact retrieval is the blueprint of facts or concept blueprint. In its basic form the blueprint is a structure of terms where each slot may hold a single term (with or without its corresponding synonyms), an N-Gram, or a Named Entity. The meaning of a particular concept of a slot can be expressed by different terms, much like the type of numerical information (date, price, or growth rate, for instance) can be expressed in different syntactic forms. Each slot is represented by a title. The titles, in turn, represent the content of the slot on a more abstract level. A blueprint, thus, consists of a hierarchy of iteratively integrated slots. Each blueprint stands for a particular concept that is further detailed by its sub-components, that is, the slots on the different levels of abstraction. Each blueprint represents not only the semantic architecture of its concept or its meaning, but also the different syntactic facets its concept may take in texts.

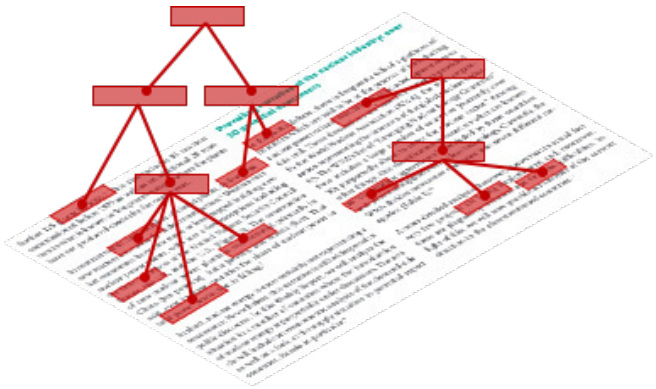

Fig. 2. Schema of a blueprint with its slots.

IV. Defining Concept Blueprints

Text Mining, in our approach, starts with seeds covering annotated definitions of basic text patterns. They include things like dates, distances, or prices. The next group of the seeds addresses proper names for locations, countries, persons, and the like. Our system design includes helper functions to detect proper names which, in general, pose a certain challenge for automatic text analysis. By applying CRF methods, we can determine these expressions. For instance, words starting with uppercase letters and immediately following special terms like “Premier ministre”, “Mr.”, or “the author” usually indicate that the following terms may be proper names of persons. Some BoWs (for countries, for instance) can also be imported from external sources.

Whenever a slot of a blueprint refers a specific term, all of its applicable synonyms need to be taken into consideration. However, not all possible synonyms are also applicable in every context. In an expression describing a certain amount of money like “to the tune of 12.65 billion U.S. dollars”, none of the synonyms of the term “tune” is applicable in this context. In some occasions, it is thus recommendable to consider the applicability of synonyms on the level of N-Grams. Thorough N-Gram analysis reveals, that expressions like the one shown in the example have synonyms like “to the amount of” or “add up to”.

Figure 3 shows the representation of a section of a text after the basic text patterns have been identified and annotated accordingly. All instances that meet the qualities of an expression representing a price are identified and marked by the blueprint `price=?"price".money.currency`. These instances are expected to be composed of an instance matching the slot (or subcomponent of the blueprint) (amount of) “money”, a further one addressing the currency and an occasional (leading or trailing) word “price” (or a synonym expression such as “at a cost of”). An optional slot is indicated by a leading question mark. Key words are stated in quotes. Internally they are mapped to their standardized (stemmed) form. Terms without quotes thus refer to the blueprint slots. The dots stand for “close proximity” which can range from “immediately adjacent” to “neighboring blueprints spread over a phrase or paragraph”.

In 2009, the UAE government commissioned Korean Electric Power Corporation (KEPCO) from South Korea to build four reactors with an output of 5.4 gigawatts (GW) at a cost of 28.2 billon U.S. dollars. This equates to a dedicated investment of 5,300 U.S. dollars per kilowatt.

<investment><time point>In <year>2009</year></time point>, the <buyer><body>UAE government</body></buyer> commissioned <seller><organization>Korean Electric Power Corporation (KEPCO)</organization></seller> from <region>South Korea</region> to build <plant>four reactors</plant> with an <output>output of <power>5.4 gigawatts (GW)</power></output> at a cost of <price>28.2 billion U.S. dollars</price></investment>. This equates to a dedicated <investment>investment of <price>5,300 U.S. dollars</price> <unit>per kilowatt</unit></investment>. No less than <price>18.7 billion U.S. dollars</price> of the total sum was financed with public money.

Fig. 3. Section of a text and its representation after basic text patterns have been identified and accordingly annotated.

Each set of slots is annotated by a title reflecting the concept or the overarching meaning of the slots. This title summarizes the content of all blueprint components on its underlying layer. It thus abstracts away the content details of the slot layer it stands for. Each such blueprint can be a slot in the next layer of abstraction.

By repeatedly applying this process the blueprint gets more layers and covers a growing text area. The blueprint then resembles a hierarchy with a general representation of covered text on its top and growing specialized representations towards its bottom.

Each single blueprint thus consists of a set of slots and its title. It forms an inseparable unit. The repeated pattern identification operates on the blueprint titles, the text sections that are so far not part of any instance of a blueprint.

V. Conclusions and Outlook

This paper presents the state of work of the design and prototypical implementation of a fact retrieval system operating on concept blueprints that can be defined by the users. It uses Named Entity Recognition and theme-specific Bag of Words to identify semantic markers in text that point to the specific meaning of a text passage.

The application areas of the content schemas are manifold. The main purpose is identifying facts in texts and representing them in a distilled and standardized way in their

respective context. This facilitates the comparison of representations of facts in different sources and, thus, supports the detection of fake news and misinformation.

Named entities and terms from BoWs identify the meaning words as they appear in a phrase or fragment of text. However, they also explicitly include numerical data that are very important for the correct reflection of meaning in text. Iteratively applying standardization to already extracted and annotated pieces of text creates semantic hierarchies which, in turn, reflect the meaning of terms in a more general or more detailed (or specified) context. This, in turn, makes text comparisons more precise and versatile.

Our approach and our prototype are still work in progress, but we already noticed that our content schemas have a certain proximity to ontologies. We use the schemas for text interpretation on a basic level and gradually produce concept hierarchies. However, we clearly see the necessity to add more functionality to schemas, in particular, when parts of the schema address factual (i.e. numerical) information. Quite often, calculations can be helpful to check the plausibility of statements based on numerical information. The standardized representation of facts enables the opportunity to include (at least some decent) inference mechanisms. The representation of extracted instances can be used to link data processing features to the slots of the blueprints.

```
<investment>
    <time_point>
        <year>
            2009
        </year>
    </time_point>
    <buyer>
        <body>
            UAE government
        </body>
    </buyer>
    <seller>
        <organization>
            Korean Electric Power Corporation (KEPCO)
        </organization>
    </seller>
    <region>
        South Korea
    </region>
    <plant>
        four reactors
    </plant>
    <output>
        <power>
            5.4 gigawatts (GW)
        </power>
    </output>
    <price>
        28.2 billion U.S. dollars
    </price>
</investment>
```

Fig. 4. Section of a representation of a fact as annotated and extracted by the blueprint “investment” (see also fig. 3).

These features can, for instance, deduce a specific date that needs to be assigned with the slot containing the word “yesterday”. Representing a text passage in machine-processable form like the one shown in figure 4 offers the opportunity to extract all information concerning investments in nuclear energy power in order to prepare it for automatic reporting.

A further objective of our approach is a stronger involvement of humans in the development and management of text mining tools, in general, to enhance the adoption of this technology on a broader scale. This involvement results in a more active role of the users in designing, controlling, and adapting the learning process that feeds, in this case here, the automatic detection of facts in text. The syntax for the definition of a blueprint is easy to learn. Even users without technical background are in the position to write definitions for concept blueprints. In the next phase of the development of our prototype, the users will be involved more closely in the training of the semi-automatic processes to detect blueprints that come semantically close to existing definitions.

REFERENCES

[1] J. L. Kolodner, “Requirements for natural language fact retrieval”,
[2] *Proceedings of the ACM '82 conference*, 1982, pp. 192–198.
[3] N. Fuhr, “Integration of probabilistic fact and text retrieval”, *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval,* 1992, pp 211–222.
[4] M. Keikha, J. H. Park, W. B. Croft, and M. Sanderson, “Retrieving Passages and Finding Answers”, *Proceedings of the 2014 Australasian Document Computing Symposium,* 2014, 81–84.
[5] N. Balasubramanian, J. Allan, and W. B. Croft, “A comparison of sentence retrieval techniques”, *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 813–814.
[6] R. W. White, S. M. Drucker, G. Marchionini, M. Hearst, and M. C. Schraefel, “Exploratory search and HCI: designing and evaluating interfaces to support exploratory search interaction”, *CHI '07 Extended Abstracts on Human Factors in Computing Systems*, 2007, pp. 2877–2880.
[7] A. T. Nguyen, A. Kharosekar, S. Krishnan, and S. Krishnan, “Believe it or not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking”, *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology,* 2018, pp. 189–199.
[8] H. E. Wynne and Z. Z. Wint, “Content Based Fake News Detection Using N-Gram Models”. *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services (iiWAS2019)*, 2019, pp. 669–673.
[9] W. A. Woods, “Context-Sensitive Parsing”. *Communications of the ACM 13(7),* 1996, pp. 413–445.
[10] F. Sha, F. and F. Pereira, F., “Shallow Parsing with Conditional Random Fields”, *Proceedings of the HLT-NAACL conference*, 2003, pp. 134-141.
[11] D. Freitag and A. McCallum, “Information Extraction with HMM Structures Learned by Stochastic Optimization”, *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, 2000, pp. 584-589.
[12] J. Cowie and W. Lehnert, “Information Extraction”. Communications of the ACM 39(1): 80–91.
[13] G. Salton, J. Allan, C. Buckely, A. Singhal, “Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts”, in: Karen Sparck Jones and Peter Willett, *Readings in Information Retrieval*, San Francisco, 1997, pp. 478–483.
[14] J. Jancsary, F. Neubarth, S. Schreitter, and H. Trost, “Towards a context-sensitive online newspaper”. *Proceedings of the 2011 Workshop on Context-awareness in Retrieval and Recommendation*, 2011, pp. 2–9.