Content and Knowledge-Based Approaches to Measure Document Similarity

UIUC CS 410 Technology Review

Daiyun Xu (web id: daiyunx2)

I. Introduction

How related are "cat" and "mouse"? And what about "preparing a manuscript" and "writing an article"? Reasoning about semantic relatedness of natural language utterances is routinely performed by humans but remains an insurmountable obstacle for computers. Humans do not judge text relatedness merely at the level of text words. Words trigger reasoning at a much deeper level that manipulates concepts—the basic units of meaning that serve humans to organize and share their knowledge. Thus, humans interpret the specific wording of a document in the much larger context of their background knowledge and experience.

In the literature, knowledge-based approaches have been employed for improving existing techniques in Natural Language Processing and Information Retrieval domains. Yet, there is much room for improvement in order to effectively exploit these rich models in these fields. For instance, in the context of inter-document similarity, which plays an important role in many NLP and IR applications, classic techniques rely solely on syntactic information and are usually based on Vector Space Models, where the documents are represented in a vector space having document words as dimensions.

Nevertheless, such techniques fail in detecting relationships among concepts in simple scenarios like the following sentences: "Fight Club's estranged narrator leaves his lackluster job when he comes under the thrall of Tyler Durden, an enigmatic young man who holds secret after-hours boxing matches in the basement of bars." and "An insomniac office worker and a devil-may-care soap maker form an underground fight club that evolves into much more."

These two sentences, abstracts of the book "Fight Club" and its movie adaptation, contain highly related concepts (e.g., "estranged narrator leaves his lackluster job" refers to "an insomniac office worker", "boxing matches in the basement of bars" refers to "underground fight club") . To find the encoded relationship between these two sentences and furthermore develop good document similarity algorithms to rank abstract text documents, we want to investigate both content-based approaches and knowledge-based approaches and see which one is more useful to compute the document similarity for my course project, an information retrieval application to recommend and rank relevant text documents from two genres, movies and books.

II. Content-based approaches

The standard document representation technique is the Vector Space Model [1]. Each document is expressed as a weighted high-dimensional vector, the dimensions corresponding to individual features such as words. The result is called the bag-of-words model and it is the first example of content-based approach. The limitation of this model is that it does not address polysemy (the same word can have multiple meanings) and synonymy (two words can represent the same concept).

Another technique belonging the content-based group is Latent Semantic Analysis (LSA) [2], which assumes that there is a latent semantic structure in the documents it analyzes. Its goal is to extract this latent semantic structure by applying dimensionality reduction to the terms-document matrix used for representing the corpus of documents.

Finally, in the context of Information Retrieval, probabilistic models are employed for ranking documents according to their relevance (similarity) to a given search query, i.e., similarities are computed as probabilities that a document representation matches or satisfies a query. Among them, the most popular are: Okapi BM-25 [3] and language modeling approaches.

III. Knowledge-enriched approaches

Recently, a lot of effort has been employed in designing new techniques for text similarity that use information contained in knowledge bases. After reviewing literature, I have gathered several promising knowledge-enriched approaches using domain knowledge like Wikipedia and DBpedia.

A. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic [4]

The author proposes Explicit Semantic Analysis (ESA) to map the documents to Wikipedia articles, and to represent each document as a vector of features extracted from both the document and the related articles text. Thus, the similarity of two documents can be computed through any vector space comparison algorithm.

They use machine learning techniques to build a semantic interpreter that maps fragments of natural language text into a weighted sequence of Wikipedia concepts ordered by their relevance to the input. Each Wikipedia concept is represented as an attribute vector of words that occur in the corresponding article. Entries of these vectors are assigned weights using TFIDF scheme. To speed up semantic interpretation, they build an inverted index, which maps each word into a list of concepts in which it appears. They also use the inverted index to discard insignificant associations between words and concepts by removing those concepts whose weights for a given word are too low.

Empirical evaluation confirms that using ESA leads to substantial improvements in computing word and text relatedness. Compared with the previous state of the art, using ESA results in notable improvements in correlation of computed relatedness scores with human judgements and is easy to explain to human users.

B. *WikiWalk: Random walks on Wikipedia for Semantic Relatedness [5]*

Another document similarity technique that leverages the information contained in Wikipedia is WikiWalk, where the personalized PageRank on Wikipedia pages is used, with a personalization vector based on the ESA weights on concepts detected in the documents, to produce a vector used for estimating the similarity.

It is based on random walk algorithms and use Wikipedia-based methods to construct the graph. they obtain a relatedness score between a pair of texts by performing random walks over a graph to compute a stationary distribution for each text. the score is simply the cosine similarity between the distributions.

They simply treat the articles as vertices, and the links between articles as the edges. In order to apply personalized PageRank to a given passage of text or word, they construct a custom teleport vector, representing the initial distribution of mass over the article nodes.

A big drawback of this approach is the computational cost; indeed, for each document they have to execute first ESA and then compute the personalized PageRank on the whole Wikipedia.

C. *Semantic Relatedness Using Salient Semantic Analysis [6]*

In this paper, authors proposed a novel unsupervised method for semantic relatedness that generates a semantic profile for words by using salient conceptual features gathered from encyclopedic knowledge. The model is built on the notion that the meaning of a word can be represented by the salient concepts found in its immediate context, by using salient concepts gathered from encyclopedic knowledge. The main idea underlying their method is that they can determine the semantic relatedness of words by measuring the distance between their concept-based profiles, where a profile consists of salient concepts occurring within contexts across a very large corpus.

They derive semantic profiles based on the Wikipedia corpus by using one of its most important properties – the linking of concepts within articles. The links available between Wikipedia articles, obtained either through manual annotation by the Wikipedia users or using an automatic annotation process, to determine the meaning and the saliency of a large number of words and phrases inside this corpus. These links are regarded as clues or salient features within the text that help define and disambiguate its context. They can measure the semantic relatedness of words by using their concept-based profiles, where a profile is constructed using the co-occurring salient concepts found within a given window size in a very large corpus.

The evaluation on standard word-to-word and text-to-text relatedness benchmarks confirms the superiority and consistency of this model. The performance of the model seems to be independent of the distance metric used in either cosine or SOCPMI evaluation. This fact provides additional support for the underlying assumption about profiling words using strong unambiguous word-concept associations.

D. *Knowledge-based Graph Document Modeling [7]*

Authors proposed a method for exploiting large amounts of machine-readable knowledge, i.e., entities and semantic relations, encoded within DBpedia, in order to provide a structured, graph-based, representation of natural language texts. Their method starts with the output of an entity disambiguator, which is used to identify a set of concepts from the input texts. Next, connecting paths between entities are collected, in order to identify the sub-graph of DBpedia covered by each document. Nodes in the semantic graph consist of concepts capturing the main topics of the documents. In addition, edges in the graph are weighted to identify the semantic relations that are most relevant for these concepts.

Finally, they view computing semantic similarity as a matching problem between the concepts of different documents, and apply a Graph Edit Distance based similarity measure to identify the ‘best’ connecting paths between the documents’ concepts.

E. *Computing inter-document similarity with Context Semantic Analysis [8]*

Authors proposed Context Semantic Analysis (CSA), a novel knowledge-based technique for estimating inter-document similarity. The technique is based on a Semantic Context Vector, which can be extracted from a knowledge base and stored as metadata of a document and employed to compute inter-document similarity.

The method CSA differs from other methods because it aims to be a general approach that can use any knowledge base expressed according to the Semantic Web standard, so that all datasets belonging to the Linked Open Data cloud (more than one thousand) can be used as source of knowledge. They focus on RDF knowledge bases: an RDF knowledge base can be considered a set of facts (statements), where each fact is a triple of the form <subject, predicate, object>. In their approach a document is represented as a vector and the similarity can be estimated more effortlessly by using cosine similarity.

They showed the consistency of CSA with respect to human judges and how it outperforms standard (i.e., syntactic) inter-document similarity methods. The method can exploit any generic RDF knowledge bases. In order to evaluate CSA they employed two generic domain knowledge bases, DBpedia and Wikidata, and they showed how it can improve the performance of an existing IR framework.

IV. CONCLUSION

Document similarity has been one the main research areas of the last years due to a wide range of its applications in tasks such as information retrieval, text classification, document clustering, topic detection, etc. In this field a lot of techniques have been proposed but I can group them in two main categories, content-based and knowledge-based approaches, where the main difference is that the first group uses only textual information contained in documents while the second one enriches these documents by extracting information from other sources, usually knowledge bases. empirically, knowledge-based techniques are more accurate than content-based techniques in their advantages of inferring underlying human knowledge and relationship from plain text.

In this review, I investigated old and new techniques in many content and knowledge-based approaches. I believe knowledge-based are more relevant to my use case. So, I reviewed more state-of-art knowledge-based models including Explicit Semantic Analysis, Random wiki-walks, Salient Semantic Analysis, Knowledge-based Graph, and Context Semantic Analysis. Except for Knowledge-based Graph, which uses DBpedia, most approaches use Wikipedia to build the knowledge graph because it is currently the largest knowledge repository on the Web. Wikipedia is available in dozens of languages, while its English version is the largest of all with 400 million words in over one million articles.

For my course project, an application to recommend and rank relevant text documents from movies and books genres, I will also consider knowledge-based approaches and use Wikipedia as the main source as most adapted story movies have their individual Wikipedia pages. Context Semantic Analysis, being the latest model, and Explicit Semantic Analysis, being the most fundamental work, will be the two strongest candidate to implement in my application.

REFERENCES

[1] Turney P.D., Pantel P. "From frequency to meaning: Vector space models of semantics" J. Artif. Int. Res., 37 (1) (2010), pp. 141-188.

[2] Dumais S.T. "Latent semantic analysis" Annu. Rev. Inf. Sci. Technol., 38 (1) (2004), pp. 188-230.

[3] Robertson S.E., Zaragoza U. "The probabilistic relevance framework: BM25 and beyond" Found. Trends Inf. Retr., 3 (4) (2009), pp. 333-389,.

[4] Gabrilovich E., Markovitch S. "Computing semantic relatedness using wikipedia-based explicit semantic analysis" Proceedings of the 20th International Joint Conference on Artifical Intelligence, IJCAI'07, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2007), pp. 1606-1611.

[5] Yeh E., Ramage D., Manning C.D., Agirre E., Soroa A. "Wikiwalk: random walks on wikipedia for semantic relatedness", Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing, Association for Computational Linguistics (2009), pp. 41-49.

[6] Hassan S., Mihalcea R. "Semantic relatedness using salient semantic analysis" Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI'11, AAAI Press (2011), pp. 884-889.

[7] Schuhmacher M., Ponzetto S.P. "Knowledge-based graph document modeling" Proceedings of the 7th ACM International Conference on Web Search and Data Mining, WSDM '14, ACM, New York, NY, USA (2014), pp. 543-552. Fabio B., Domenico B., Sonia B., Giovanni S., "Computing inter-document similarity with Context Semantic Analysis" Inf. Sys., 80 (2019), pp. 136-147,.

[8] Fabio B., Domenico B., Sonia B., Giovanni S., "Computing inter-document similarity with Context Semantic Analysis" Inf. Sys., 80 (2019), pp. 136-147,.