Toward Document Understanding for Information Retrieval

Mostafa Dehghani
University of Amsterdam
dehghani@uva.nl

Abstract

Document understanding for the purpose of assessing the relevance of a document or passage to a query based only on the document content appears to be a familiar goal for information retrieval community, however, this problem has remained largely intractable, despite repeated attacks over many years. This is while people are able to assess the relevance quite well, though unfamiliar topics and complex documents can defeat them. This assessment may require the ability to understand language, images, document structure, videos, audio, and functional elements. In turn, understanding of these elements is built on background information about the world, such as human behavior patterns, and even more fundamental truths such as the existence of time, space, and people. All this comes naturally to people, but not to computers!

Recently, large-scale machine learning has altered the landscape. Deep learning has greatly advanced machine understanding of images and language. Since document and query understanding incorporate these elements, deep learning can hold great promise. But it comes with a drawback: general purpose representations (like CNNs for images) have proved somewhat elusive for text. In particular, embeddings act as a distributed representation not just of semantic information but also application-specific learning, which are hard to transfer. In short, conditions seem right for a renewed attempt on the fundamental document understanding problem.

1 What is document understanding?

In order to think about the way we can approach this problem, I think we should first answer some questions: Can we understand documents? What is "understanding"? Getting at the true meaning of a document? Okay, but then what is "meaning"? How do we even approach such an ill-defined goal?

As an initial disclaimer, the goal here is not to understand documents in some abstract sense, but rather to understand documents well enough to perform a particular task: determining whether the document is relevant to a query. This assumption simplifies the problem. In other words, we dont need to understand a calculus tutorial well enough to compute integrals; rather, only well enough to determine whether it is relevant for queries such as "integrate rational functions". Of course, we can hope that pursuing this narrow goal will produce insights that will someday allow us to understand documents in a broader sense, but even for this narrower goal, specific problem instances seem to be extraordinarily difficult, with a lot of challenges. As an example of such a challenge, how can we map a long natural-language text into a format suitable for large-scale machine learning without destroying useful information? For example, if we supply the text to a deep neural network as a bag of tokens, we discard both low-level information (the characters that make up the tokens) and higher-level information (word order) before deep learning enters the picture.

2 How can a search engine successfully fake document understanding?

I am not aware of what search engines do, but I can imagine how they can successfully fake document understanding. Consider a search engine as a person who works in a library full of Persian books, but he does not read or speak this language. A visitor asks for a book about, say, " ". The search engine has no idea what this means. But he searches the shelves for books with those two words, particularly in prominent places such as titles or chapter names. He places let's say 10 candidates on the counter. The visitor frowns at a few, but then smiles at another and takes that one to read. Later, when someone else asks for a book about " ", he shows the smile-inducing book and holds back the frown-inducing ones. After employing this strategy for some years, he becomes a pretty good recommender of Persian books, despite having no clue what any of them actually say. Of course, he might mishandle complex or new topics. The key idea is that he substitutes an ability to interpret and remember human reactions for the ability to read the text himself1.

This can be actually the case as search engines just want to provide users with documents relevant to their queries, even though they cannot understand the documents themselves. To achieve this goal, a search engine can employ very simple techniques, e.g. using keyword-oriented heuristics like ensuring that all important queries terms (or synonyms) appear in the document and prefer documents where these terms appear often, prominently, close to one another, etc. plus favoring documents to which people react positively, with respect to the evidence like clicks.

3 What is wrong with fake document understanding?

In order to fake understanding of documents, search engines recall human reactions to those documents. The form of reaction they mostly rely upon in web ranking is clicks on search results. This user feedback can be extremely helpful and central to the strategy I explained before, but has problems like:

People sometimes click on engaging or higher-ranked results, but with irrelevant content. Clicks are based on limited information. Users click on documents before reading them so user clicks are perhaps driven more strongly by the presentation of results on the search page, than by the content of the underlying documents. Fresh content lacks clicks. So although click-through information is probably the best form of user feedback that can be used for document understanding, but has serious drawbacks. In particular, understanding documents to explain user clicks is different from understanding documents for the purpose of assessing relevance. In machine learning terms, the training data does not precisely match the task.

4 How can we benefit from real document understanding?

Although faking document understanding might look sufficient, but there are some particular points that we can benefit from real document understanding including but not limited to:

Getting benefit in ranking around fresh, recently-changed, and rarely-seen documents. Improving performance on verbose natural language queries, which may be more common in communications with an assistant. Classic relevance assessment mechanisms perform relatively poor in these areas. Seeing documents from a human perspective which can lead to new features and even new applications. For instance, we can presumably save people time by presenting results in a manner that makes clear the case for relevance. From Memorization to Generalization I think the first step for moving from fake document understanding to the real one would be moving from pure memorization toward generalization. A search engine might become a good Persian librarian purely by memorizing which books people like and dislike on each topic. But he can do even better by noting patterns in the behaviors he observes. For example, he might notice that books in a particular series always elicit a frown, no matter what the library visitor specifically requested, so he should probably stop suggesting those. Following this, a search engine can memorize past user behavior in the hope of predicting future behavior. Going beyond this, a search engine can generalize; that is, find patterns in past user behavior that help to anticipate behavior in new situations. In other words, a search engine can learn not only highly-specific rules in form of "show document D for query Q", but also rules that are more widely applicable.

As I mentioned, to start we can translate "understanding" as "memorizing with the ability of generalization" which is explaining user behavior in terms of stimuli to which people actually respond. In this light, we are currently well short of understanding, but the goal is not hopelessly out of reach.

5 Useful document representations

When document understanding for information retrieval comes to action, in the first step, we can think of it as "Learning Useful Document Representations for Search". Generating representation for documents is still one the key challenges in information retrieval. So we want to process the document in order to produce a representation of it that preserves our ability to judge relevance while stripping away nonessential data.

Most of the time, due to efficiency reasons, this is done as a preprocessing stage where we generate query-independent document representation and it has one straightforward motivation: processing documents online would be terribly expensive. Note that this is not to suggest that query-dependent document representations are not useful, for instance, we can train a neural network which predicts the relevance of a query and a document based on words near hits which could be both efficient and effective.

Since deep learning turned to be a tool using which we can renew our attempts for representing documents, it is good to think about this question again that "What does make for useful document representations?”. Of course representations that lead to better results in search are appreciated, but it is not enough. Here, I’m going to shortly talk about some characteristics that I think a useful query-independent document representation for the task of search should possess.

Good representations of documents should ideally satisfy the following properties for maximal utility in information retrieval:

1. **Semantic**: The representation should be the same or similar even if the text is rewritten in a different manner as long as it has the same meaning. In other words, it should be resilient to paraphrases.
2. **Similarity/Relevance measure**: A similarity measure between two document representations should be computable. This would in a sense be the inverse of a distance measure1. Alternatively, the similarity between a query representation and a document representation can be considered, which is a measure of whether the intent of query is covered by the document.
3. **Composability**: Composability allows us to combine representations in various ways, for instance, compose representations at the sentence level to get a representation for a paragraph. We can further compose those to get a representation of a document. This way we consider documents as volumes rather than points. As another application of composability, we can think of generating query representations based on its (pseudo-)relevant documents, by performing pseudo-relevance feedback either online or offline. It is particularly nice to come up with representations where compositions can be done using the linear combination, as it is simple to understand and play with.
4. **Similarity is congruent with relevance to queries**: In other words, if we have relevance judgments for results of a query, the level of similarity between relevant results is higher than the level of similarity between irrelevant results and irrelevant results are dissimilar to relevant results.

There are also some additional optional properties that are nice to have:

5. **Common representation for queries and documents**: If queries are representable in the same space as documents, then a similarity measure between the query representation and the document representation could be indicative of relevance, although this is hard to satisfy in practice if a point representation is used. The reason this property is not necessary is that in practice, we could derive a query representation from properties (3) and (4): For example, we could represent each query by a suitable linear combination of the representations of the search results from a first pass, or we can compute offline representations of past search queries using a similar trick and try to predict the representation for a new user query by learning from representations of similar past search queries. Note that conversely, we could start with query representations and derive a document representation as well. However, a document is much richer in terms of directly available content signals than a query, so it seems reasonable to expect more useful representations for both documents and queries by starting from document representations.
6. **Transparency**: The representation should be transparent and debuggable. I know that effective embedding representations learned using neural-based models might not be as understandable, but having insight and a handle to prob the representations

would help a lot to improve them. Hierarchical: Semantics for larger and larger blocks of text might be best represented hierarchically. When writing formally, the title serves as a high-level summary. Each paragraph should be self-coherent and support the root concept. Lastly, each sentence should have a single "thought." When comparing queries to documents, different granularity might serve different needs. For example, a factoid queries need to find a short trustworthy passage (maybe even a single sentence). The overall topic of the page is related, but less useful than determining if the query is satisfied (has the answer). On the other side of the spectrum, broad information seeking (like exploratory search) probably cares more about the overall topic of the page, than each support sentence.

7. **Language agnostic**: If multiple documents have the exact same semantics but are in different languages, they should probably have a similar representation. Represents more than just natural language: In web search, where documents are more than text, i.e. they have structure, they include images, etc. we should be able to cover them with a similar representation.

6 Conclusion

Fully understanding documents is probably far out of reach. For example, coping with humor and complex inferences would probably require true artificial intelligence. So, to some extent, we must continue to "fake" document understanding through memorization but try to increase the generalization for a long time to come. But we also need to keep in mind that we need to move toward a direction in order to reduce our reliance on fake document understanding!