Advancing Natural Language Understanding with Collaboratively Generated Content

EVGENIY GABRILOVICH

Yahoo! Research

Proliferation of ubiquitous access to the Internet enables millions of Web users to collaborate online in a variety of activities. Many of these activities result in the construction of large repositories of knowledge, either as their primary aim (e.g., Wikipedia) or as a by-product (e.g., Yahoo! Answers). In this paper, we discuss how to use the cornucopia of world knowledge encoded in the repositories of collaboratively generated content (CGC) for advancing computers' ability to process human language.

Prior to the advent of CGC repositories, many computational approaches to natural language employed the WordNet electronic dictionary (Fellbaum, 1998), which covers approximately 150 thousand words painstakingly encoded by professional linguists over the course of more than 20 years. In contrast, the collaborative Wiktionary project (www.wiktionary.org) includes more than 2.5 million words in English alone. Encyclopaedia Britannica published since 1798 (sic!) has approximately 65 thousand articles, while Wikipedia has over 3.7 million articles in English and over 15 million articles in over 200 other languages. Ramakrishnan and Tomkins (2007) estimated the amount of user-generated content produced worldwide on a daily basis to be 8-10 Gigabytes, and this amount has likely increased considerably since then.

Repositories of collaboratively generated content as an enabling resource

The unprecedented amounts of information in CGC enable new, knowledge-rich approaches to natural language processing, which are significantly more powerful than the conventional word-based methods. Considerable progress has been made in this direction over the last few years. Examples include explicit manipulation of human-defined concepts and their use to augment the bag of words in information retrieval (Egozi et al., 2011), or using Wikipedia for better word sense disambiguation (Bunescu and Pasca, 2006; Cucerzan, 2007).

One way to use CGC repositories is to treat them as huge additional corpora, for instance, to compute more reliable term statistics or to construct comprehensive lexicons and gazetteers. They can also be used to extend existing knowledge repositories, increasing the concept coverage and adding usage examples for previously listed concepts. Some CGC repositories, such as Wikipedia, record each and every change to their content, thus making the document authoring process directly observable. This abundance of editing information allows us to come up with better models of term importance in

documents, assuming that terms introduced earlier in the document life are more central to its topic. The recently proposed Revision History Analysis (Aji et al., 2010) captures this intuition to provide more accurate retrieval of versioned documents.

An even more promising research direction, however, is to distill the world knowledge from the structure and content of CGC repositories. This knowledge can give rise to new representations of texts beyond the conventional bag of words, and allow reasoning about the meaning of texts at the level of concepts rather than individual words or phrases. Consider, for example, the following text fragment: "Wal-Mart supply chain goes real time". Without relying on large amounts of external knowledge, it would be quite difficult for a computer to understand the meaning of this sentence. Explicit Semantic Analysis (Gabrilovich and Markovitch, 2009) offers a way to consult Wikipedia in order to fetch highly relevant concepts such as "Sam Walton" (the Wal-Mart founder); "Sears", "Target" and "Albertsons" (prominent competitors of Wal-Mart); "United Food and Commercial Workers" (a labor union that has been trying to organize Wal-Mart's workers); "Hypermarket" and "Chain store" (relevant general concepts). Arguably, the most insightful concept generated by consulting Wikipedia is "RFID" (Radio Frequency Identification), a technology extensively used by Wal-Mart to manage its stock. None of these concepts are explicitly mentioned in the given text fragment, yet when available they help shed light on the meaning of this short text.

In the remainder of this article, we first discuss using CGC repositories for computing semantic relatedness of words, and then proceed to higher-level applications such as information retrieval.

Computing semantic similarity of words and texts

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 challenging 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.

Prior work on semantic relatedness was based on purely statistical techniques that did not make use of background knowledge (Deerwester et al., 1990), or on lexical resources that incorporate limited knowledge about the world (Budanitsky and Hirst, 2006). CGC-based approaches differ from the former in that they manipulate concepts explicitly defined by humans, and from the latter – in the sheer number of concepts and the amount of background knowledge. One class of new approaches to computing semantic relatedness uses the structure of CGC repositories, such as category hierarchies (Strube and Ponzetto, 2006) or links among the concepts (Milne and Witten, 2008). Given a pair of words whose relatedness needs to be assessed, these methods map them to relevant concepts (e.g., articles in Wikipedia), and then use the structure of the repository to compute the relatedness between these concepts. Gabrilovich and Markovitch (2009) proposed an alternative approach that uses the entire content of Wikipedia, and represents the meaning of words and texts in the space of Wikipedia

concepts. Their method – called Explicit Semantic Analysis (ESA) – represents texts as weighted vectors of concepts. The meaning of a text fragment is thus interpreted in terms of its affinity with a host of Wikipedia concepts. Computing semantic relatedness of texts then amounts to comparing their vectors in the space defined by the concepts, for example, using the cosine metric.

Subsequently proposed approaches offer ways to combine the structure-based and concept-based methods in a principled manner (Yeh et al., 2009). Beyond Wikipedia, Zesch et al. (2008) proposed a method for computing semantic relatedness of words using Wiktionary. Recently, Radinsky et al. (2011) proposed a way to augment the knowledge extracted from CGC repositories with temporal information, by studying patterns of word usage over time. Consider, for example, an archive of The New York Times spanning 150 years. Two words such as "war" and "peace" might rarely co-occur in the same articles, yet their patterns of use over time might be similar, which allows us to better judge their true relatedness.

Concept-based information retrieval

Information retrieval systems traditionally rely on textual keywords to index and retrieve documents. Keyword-based retrieval may return inaccurate and incomplete results when different keywords are used to describe the same concept in the documents and in the queries. Furthermore, the relationship between those related keywords may be semantic rather than syntactic, and capturing it thus requires access to comprehensive human world knowledge. Previous approaches have attempted to tackle these difficulties by using manually-built thesauri, by relying on term co-occurrence data, or by extracting latent word relationships and concepts from a corpus. Explicit Semantic Analysis introduced in the previous section, which represents the meaning of texts in a very high-dimensional space of Wikipedia concepts, has been shown to offer superior performance over the previous state-of-the-art algorithms. In contrast to the task of computing semantic relatedness, which usually deals with short texts whose overlap is often empty, information retrieval usually deals with longer documents. It is noteworthy that in such cases optimal results can be obtained by extending the bag of words with concepts, rather than merely relying on the conceptual representation alone.

Intuitively, one might expect domain-specific knowledge to be key for processing texts in terminology-rich domains such as medicine. However, as Gabrilovich and Markovitch (2007) showed, it is the general purpose knowledge that leads to much higher improvements in text classification accuracy. In the follow-up article (Gabrilovich and Markovitch, 2009), the authors also showed that using larger repositories of knowledge (e.g., later Wikipedia snapshots) leads to superior performance as more knowledge becomes available.

Potthast et al. (2008) and Sorg and Cimiano (2008) independently proposed CL-ESA, a cross-lingual extension to Explicit Semantic Analysis. Using cross-language links available between a growing number of Wikipedia articles, the approach allows to map the meaning of texts across different languages. This allows, for example, to formulate a query in one language and then use it to retrieve documents written in a different language.

Conclusion

Publicly available repositories of collaboratively generated content encode massive amounts of human knowledge about the world. In this paper, we showed that the structure and content of these repositories can be used to augment representation of natural language texts with information that cannot be deduced from the input text alone.

Using knowledge from CGC repositories leads to double-digit accuracy improvements in a range of tasks, from computing semantic relatedness of words and texts to information retrieval and text classification. The most important aspects of using exogenous knowledge are its ability to addresses synonymy and polysemy, which are arguably the two most important problems in natural language processing. The former manifests itself when two texts discuss the same topic using different words, and the conventional bag-of-words representation is not able to identify this commonality. On the other hand, the mere fact that the two texts contain the same polysemous word does not necessarily imply that they discuss the same topic, since that word could be used in the two texts in two different meanings. We believe that concept-based representations are so successful because they allow generalizations and refinements, which partially address synonymy and polysemy.

References

Aji, A., Y. Wang, E. Agichtein, and E. Gabrilovich. 2010. Using the past to score the present: Extending term weighting models with Revision History Analysis. Proceedings of the 19th ACM Conference on Information and Knowledge Management.

Budanitsky, A., and G. Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. Computational Linguistics 32(1):13-47.

Bunescu, R., and M. Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. Pp. 9-16 in Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics.

Cucerzan, S. 2007. Large-scale named entity disambiguation based on Wikipedia data. Pp. 708-716 in Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.

Deerwester, S., S. Dumais, G. Furnas, T. Landauer, and R. Harshman, 1990. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science 41(6):391-407.

Egozi, O., E. Gabrilovich, and S. Markovitch. 2011. Concept-based information retrieval using Explicit Semantic Analysis. ACM Transactions on Information Systems 29(2).

Fellbaum, C. 1998. WordNet: An electronic lexical database. MIT Press, Cambridge, MA.

Gabrilovich, E., and S. Markovitch, 2007. Harnessing the Expertise of 70,000 Human Editors: Knowledge-Based Feature Generation for Text Categorization. Journal of Machine Learning Research 8:2297-2345.

Gabrilovich, E., and S. Markovitch. 2009. Wikipedia-based semantic interpretation for natural language processing. Journal of Artificial Intelligence Research 34:443-498.

Milne, D., and I.A. Witten, 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. Proceedings of the 2008 AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy.

Potthast, M., B. Stein, and M. Anderka. 2008. A Wikipedia-based multilingual retrieval model. Proceedings of the European Conference on Information Retrieval.

Radinsky, K., A. Agichtein, E. Gabrilovich, and S. Markovitch. 2011. A Word at a Time: Computing word relatedness using Temporal Semantic Analysis. Proceedings of the 20th International World Wide Web Conference.

Ramakrishnan, R., and A. Tomkins. 2007. Toward a PeopleWeb. IEEE Computer 40(8):63-72

Sorg, P., and P. Cimiano. 2008. Cross-lingual information retrieval with Explicit Semantic Analysis. Working Notes for the CLEF Workshop.

Strube, M., and S.P. Ponzetto. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. Pp. 1419-1424 in Proceedings of the 21st National Conference on Artificial Intelligence.

Yeh, E., D. Ramage, C.D. Manning, E. Agirre, and A. Soroa. 2009. WikiWalk: Random walks on Wikipedia for semantic relatedness. Pp. 41-49 in Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing.

Zesch, T., C. Mueller, and I. Gurevych. 2008. Using Wiktionary for computing semantic relatedness. Pp. 861-866 in Proceedings of the 23rd AAAI Conference on Artificial Intelligence.