Analysis of Document Diversity through Sentence-Level Opinion and Relation Extraction

Alessandro Moschitti

Department of Computer Science and Information Engineering
University of Trento
Via Sommarive 14, 38100 POVO (TN) - Italy
moschitti@disi.unitn.it

Abstract. Diversity in document retrieval has been mainly approached as a classical statistical problem, where the typical optimization function aims at diversifying the retrieval items represented by means of language models. Although this is an essential step for the development of effective approaches to capture diversity, it is clearly not sufficient. The effort in Novelty Detection has shown that sentence-level analysis is a promising research direction. However, models and theory are needed for understanding the difference in content of the target sentences.
In this paper, an argument for using current state-of-the-art in Relation and Opinion Extraction at the sentence level is made. After presenting some ideas for the use of the above technology for document retrieval, advanced extraction models are briefly described.

Keywords: Relation Extraction; Opinion Mining; Diversity in Retrieval

1 Introduction

Diversity in document retrieval has been mainly approached as a classical statistical problem, where the typical optimization function aims at diversifying the retrieval items represented by means of language models, see for example the novelty detection track [2]. Although, this is an essential step for the development of effective approaches to diversity in retrieval, it is not sufficient. Indeed, while for standard document retrieval, frequency counts and the related weighting schemes help in defining the most probable user information needs, they play an adversary role in capturing diversity.

For example, when retrieving documents related to the entity *Michael Jordan*, a huge amount of text will be related to the basket player; perhaps other items will be related to the Jordan, statisticians and professor, but very few of them, e.g., will be devoted to the *Michael Jordan* accounting employee for Rolfe, Benson LLP. The occurrences of the latter in Web documents will be so small that no powerful language model will be able to effectively exploit them, considering the ocean of the basket player related information. In other words, there will not be enough statistical evidence to build a language model for such employee, consequently the related context, e.g. words, can be confused with the one of other documents unrelated to *Michael Jordan*.

The solution of this problem requires the use of techniques for fine grained analysis of document semantics. In a statistical framework this means that we need to extract features semantically related[1] to the object about which the users expressed their information needs. Such features cannot be just constituted by simple context words as the frequency problem highlighted above would prevent them to be effective. In contrast, textual relations between entities like those defined in ACE [8] provide an interesting level of characterization of the target entity. For example, the sole relation *Is_employed_at* can easily diversifies the three *Michael Jordan* above. A search engine aiming at providing diversity in retrieval will need to integrate such technology in the classical language model.

Another interesting dimension of document diversity is the opinion expressed in text. Documents can be 99% similar according to scalar product based on weighting schemes (especially if traditional stoplists are applied) but express a completely different viewpoint. This is manly due to the fact that documents reporting different opinions on some events describe them by manly only changing adjectives, adverbs and syntactic constructions. Typical opinion polarity classifiers can help to separate diverse retrieved documents but, when several events are described, the opinion analysis at the document level is ineffective. In contrast, by extracting topics, opinion holders and opinion expressions would make it possible to retrieve documents that are diverse with respect to events and opinion on them. In this perspective, one main goal of the LivingKnowledge project[2] is to reveal and analyze the diversity of the information in the Web, as well as the potential bias existing on the related sources.

In the reminder of this paper, Section 2 will report on latest results of sentence-level Relation Extraction, Section 3 will describe our approach to opinion mining in LivingKnowledge and finally, Section 4 will derive the conclusions.

2 Sentence-Level Relaton Extraction

The extraction of relational data, e.g. relational facts, or world knowledge from text, e.g. from the Web [26], has drawn its popularity from its potential applications in a broad range of tasks. The Relation Extraction (RE) is defined in ACE as the task of finding relevant semantic relations between pairs of entities in texts. Figure 1 shows part of a document from ACE 2004 corpus, a collection of news articles.

In the text, the relation between *president* and *NBC's entertainment division* describes the relationship between the first entity (person) and the second (organization) where the person holds a managerial position.

To identify such semantic relations using machine learning, three settings have been applied, namely supervised methods, e.g. [27, 7, 12, 30], semi-supervised methods, e.g. [4, 1], and unsupervised methods, e.g. [9, 3]. Work on supervised

[1] At a higher level than the simple lexical co-occurences.

[2] `http://livingknowledge-project.eu/`

> Jeff Zucker, the longtime executive producer of NBC's ”Today” program, will be named Friday as the new **president** of **NBC's entertainment division**, replacing Garth Ancier, NBC executives said.

Fig. 1. A document from ACE 2004 with all entity mentions in bold.

Relation Extraction has mostly employed kernel-based approaches, e.g. [27, 7, 5, 28, 6, 21, 29]. However, such approaches can be applied to few relation types thus distant supervised learning [14] was introduced to tackle such problem. Another solution proposed in [23] was to adapt models trained in one domain to other text domains.

Although, the supervised models are far more accurate than unsupervised approaches, they require labeled data and tend to be domain-dependent as different domains involve different relations. This is a clear limitation for the purpose of improving diversity retrieval since document aspects like entities and events are typically very diverse and thus require different sources of annotated data.

The drawback above can be alleviated by applying a form of weakly supervision, specifically named distant supervision (DS), using Wikipedia data [3, 14, 10]. The main idea is to exploit (i) relation repositories, e.g. the *Infobox*, x, of Wikipedia to define a set of relation types $RT(x)$ and (ii) the text of the page associated with x to produce the training sentences, which are supposed to express instances of $RT(x)$.

Previous work has applied DS to RE at *corpus level*, e.g., [3, 14]: relation extractors are (i) learned using such not completely accurate data and (ii) applied to extract relation instances from the whole corpus. The multiple pieces of evidence for each relation instance are then exploited to recover from errors of the automatic extractors. Additionally, a recent approach, i.e., [10], has shown that DS can be also applied at level of Wikipedia article: given a target *Infobox* template, all its attributes[3] can be extracted from a given document matching such template.

In contrast, sentence-level RE (SLRE) has been only modeled with the traditional supervised approach, e.g., using the data manually annotated in ACE [7, 12, 30, 5, 28, 29, 6, 21]. The resulting extractors are very valuable as they find rare relation instances that might be expressed in only one document. For example, the relation *President(Barrack Obama, United States)* can be extracted from thousands of documents thus there is a large chance of acquiring it. In contrast, *President(Eneko Agirre, SIGLEX)* is probably expressed in very few documents (if not just one sentence), increasing the complexity for obtaining it.

[3] This is a simpler tasks as one of the two entity is fixed.

2.1 Automated Extraction of General Purpose Relationships

We have proposed a substantial enhancements of SLRE: first, the use of DS, where the relation providers are external repositories, e.g., YAGO [24], and the training instances are gathered from Freebase [13]. These allow for potentially obtaining larger training data and many more relations, defined in different sources.

Second, we have adapted state-of-the-art models for ACE RE, based on Support Vector Machines (SVMs) and kernel methods (KM), to Wikipedia. We used tree and sequence kernels that can exploit structural information and interdependencies among possible labels. The comparative experiments show that our models are flexible and robust to Web documents as we achieve the interesting F1 of 74.29% on 52 YAGO relations. To give a very rough idea of the importance of the results, the document-level attribute extraction based on DS showed an F1 of 61% [10].

Third, we have verified the quality of our SLRE, by manually mapping relations from YAGO to ACE based on their descriptions. We designed a joint RE model combining DS and ACE data and tested it on ACE annotations (thus according to expert linguistic annotators). The improvement of 2.29 percent points (76.23%-73.94%) shows that our DS data is consistent and valuable.

Finally, since our aim is to produce RE for real-world applications, we have experimented with end-to-end systems. For this purpose, we also exploit Freebase for creating training data for our robust Named Entity Recognizer (NER). Consequently, our RE system is applicable to any document/sentence. The satisfactory F1 of 67% for the 52 YAGO relations suggests that our technology can be applied to real scenarios. This is an important piece of evidence that the use of general purpose RE technology for achieving diversity in retrieval is a viable research direction.

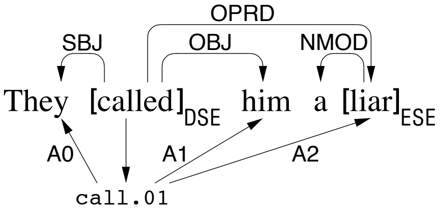

Fig. 2. Syntactic and shallow semantic structure.

3 Sentence-Level Opinion Extraction

Judgements, assessments and opinions play a crucial role in many areas of our societies, including politics and economics. They reflect knowledge diversity in perspective and goals. The vision inspiring LivingKnowledge (LK) is to consider diversity as an asset and to make it traceable, understandable and exploitable, with the goal to improve navigation and search in very large multimodal datasets (e.g., the Web itself).

To design systems that are capable of automatically analyzing opinions in *free text*, it is necessary to consider syntactic/semantic structures of natural language expressed in the target documents. Although several sources of information and knowledge are considered in LK, we here illustrate an example only focused on text. Given a natural language sentence like for example:

They called him a liar.

the opinion analysis requires to determine: (i) the opinion holder, i.e. *They*, (ii) the direct subjective expressions (DSEs), which are explicit mentions of opinion, i.e. *called*, and (iii) the expressive subjective elements (ESEs), which signal the attitude of the speakers by means of the words they choose, i.e. *liar*.

In order to automatically extract such data, the overall sentence semantics must be considered. In turn, this can be derived by representing the syntactic and shallow semantic dependencies between sentence words. Figure 2 shows a graph representation, which can be automatically generated by off-the-shelf syntactic/semantic parsers, e.g. [11], [15]. The oriented arcs, above the sentences, represent syntactic dependencies whereas the arcs below are shallow semantic (or semantic role) annotations. For example, the predicate *called*, which is an instance of the PropBank [22] frame `call.01`, has three semantic arguments: the Agent (A0), the Theme (A1), and a second predicate (A2), which are realized on the surface-syntactic level as a subject, a direct object, and an object predicative complement, respectively.

Once the richer representation above is available, we need to encode it in the learning algorithm, which will be applied to learn the functionality (subjective expression segmentation and recognition) of the target system module, i.e. the opinion recognizer. Since such graphs are essentially trees, we exploit the ability of tree kernels [16, 20, 17, 19, 18] to represent them in terms of subtrees, i.e. each subtree will be generated as an individual feature of the huge space of substructures.

Regarding practical design, kernels for structures such us trees, sequences and sets of them are available in the SVM-Light-TK toolkit (`http://disi.unitn.it/moschitti/Tree-Kernel.htm`). This encodes several structural kernels in Support Vector Machines, which is one of the most accurate learning algorithm [25].

Our initial test on the LivingKnowledge tasks suggests that kernel methods and machine learning are an effective approach to model the complex semantic phenomena of natural language.

4 Conclusion

In this paper, we have described some limits of only using language models for diversity in document retrieval. As shown by previous work in novelty detection, an analysis of document at sentence level should be carried out. In this respect, we have shown state-of-the-art natural language processing techniques for Relation Extraction and Opinion Mining, where for the former innovative approaches based on distant supervision allow for training many general purpose relation extractors.

Once accurate sentence analysis is available, several scenarios in the field of Information Retrieval open up:

– Search engines for people retrieval: the availability of automatically derived relations allows for an accurate entity disambiguation;
– Retrieval based on diversity of events: relations along with temporal information constitute basic events and are building blocks of more complex ones;
– Retrieval based on diversity in opinion: retrieval of review fragments targeting a special product or its subpart.

The FET (future emerging technology) project, LivingKnowledge, is studying such innovative approaches to diversity, although the rapid development of the above-mentioned technology suggests that such futuristic approaches are already our present.

Acknowledgements

This research has been supported by the EC project, EternalS – "Trustworthy Eternal Systems via Evolving Software, Data and Knowledge" (project number FP7 247758) and by the EC Project, LivingKnowledge – "Facts, Opinions and Bias" in Time (project number FP7 231126).

References

1. Agichtein, E., Gravano, L.: Snowball: Extracting relations from large plain-text collections. In: Proceedings of the 5th ACM International Conference on Digital Libraries. pp. 85–94 (2000)
2. Allan, J., Wade, C., Bolivar, A.: Retrieval and novelty detection at the sentence level. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. pp. 314–321. SIGIR '03, ACM, New York, NY, USA (2003)
3. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: Proceedings of IJCAI. pp. 2670–2676 (2007)
4. Brin, S.: Extracting patterns and relations from world wide web. In: Proceedings of WebDB Workshop at 6th International Conference on Extending Database Technology. pp. 172–183 (1998)
5. Bunescu, R., Mooney, R.: A shortest path dependency kernel for relation extraction. In: Proceedings of HLT and EMNLP. pp. 724–731. Vancouver, British Columbia, Canada (October 2005)
6. Bunescu, R.C.: Learning to extract relations from the web using minimal supervision. In: Proceedings of ACL (2007)
7. Culotta, A., Sorensen, J.: Dependency tree kernels for relation extraction. In: Proceedings of ACL. pp. 423–429. Barcelona, Spain (July 2004)

8. Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R.: The automatic content extraction (ace) programtasks, data, and evaluation. In: Proceedings of LREC. pp. 837–840. Barcelona, Spain (2004)
9. Hasegawa, T., Sekine, S., Grishman, R.: Discovering relations among named entities from large corpora. In: Proceedings of ACL. pp. 415–422. Barcelona, Spain (July 2004)
10. Hoffmann, R., Zhang, C., Weld, D.S.: Learning 5000 relational extractors. In: Proceedings of ACL. pp. 286–295. Uppsala, Sweden (July 2010)
11. Johansson, R., Nugues, P.: Dependency-based syntactic–semantic analysis with PropBank and NomBank. In: Proceedings of the Shared Task Session of CoNLL-2008 (2008)
12. Kambhatla, N.: Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In: The Companion Volume to the Proceedings of ACL. pp. 178–181. Barcelona, Spain (July 2004)
13. Metaweb Technologies: Freebase wikipedia extraction (wex) (March 2010), `http://download.freebase.com/wex/`
14. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of ACL-AFNLP. pp. 1003–1011. Suntec, Singapore (August 2009)
15. Moschitti, A., Coppola, B., Giuglea, A., Basili, R.: Hierarchical semantic role labeling. In: CoNLL 2005 shared task (2005)
16. Moschitti, A.: Efficient convolution kernels for dependency and constituent syntactic trees. In: Proceedings of ECML'06. pp. 318–329 (2006)
17. Moschitti, A.: Making tree kernels practical for natural language learning. In: Proccedings of EACL'06 (2006)
18. Moschitti, A.: Kernel methods, syntax and semantics for relational text categorization. In: Proceeding of CIKM 2008 (2008)
19. Moschitti, A., Quarteroni, S., Basili, R., Manandhar, S.: Exploiting syntactic and shallow semantic kernels for question/answer classification. In: Proceedings of ACL'07 (2007)
20. Moschitti, A., Zanzotto, F.M.: Fast and effective kernels for relational learning from texts. In: ICML'07 (2007)
21. Nguyen, T.V.T., Moschitti, A., Riccardi, G.: Convolution kernels on constituent, dependency and sequential structures for relation extraction. In: Proceedings of EMNLP. pp. 1378–1387. Singapore (August 2009)
22. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: An annotated corpus of semantic roles. Comput. Linguist. 31(1), 71–106 (2005)
23. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In: Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science, vol. 6323, pp. 148–163. Springer Berlin / Heidelberg (2010)
24. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago - a core of semantic knowledge. In: 16th international World Wide Web conference. pp. 697–706 (2007)
25. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience (1998)
26. Yates, A.: Extracting world knowledge from the web. IEEE Computer 42(6), 94–97 (June 2009)
27. Zelenko, D., Aone, C., Richardella, A.: Kernel methods for relation extraction. In: Proceedings of EMNLP-ACL. pp. 181–201 (2002)
28. Zhang, M., Su, J., Wang, D., Zhou, G., Tan, C.L.: Discovering relations between named entities from a large raw corpus using tree similarity-based clustering. In:

Proceedings of IJCNLP'2005, Lecture Notes in Computer Science (LNCS 3651). pp. 378–389. Jeju Island, South Korea (2005)

29. Zhang, M., Zhang, J., Su, J., , Zhou, G.: A composite kernel to extract relations between entities with both flat and structured features. In: Proceedings of COLING-ACL 2006. pp. 825–832 (2006)
30. Zhou, G., Su, J., Zhang, J., , Zhang, M.: Exploring various knowledge in relation extraction. In: Proceedings of ACL. pp. 427–434. Ann Arbor, USA (June 2005)