Crowdsourced Corpus with Entity Salience Annotations

Milan Dojchinovski[1], Dinesh Reddy[2], Tomáš Kliegr[3], Tomas Vitvar[1], Harald Sack[2]
[1] Web Intelligence Research Group
Faculty of Information Technology, Czech Technical University in Prague
firstname.lastname@fit.cvut.cz

Semantic Technologies Group
[2] Hasso Plattner Institute, University of Potsdam, Germany
firstname.lastname@hpi.de

[3] Department of Information and Knowledge Engineering
Faculty of Informatics and Statistics, University of Economics, Prague
firstname.lastname@vse.cz

Abstract

In this paper, we present a crowdsourced dataset which adds entity salience (importance) annotations to the Reuters-128 dataset, which is subset of Reuters-21578. The dataset is distributed under a free license and publish in the NLP Interchange Format, which fosters interoperability and re-use. We show the potential of the dataset on the task of learning an entity salience classifier and report on the results from several experiments.

Keywords: document aboutness, named entities, entity salience, entity importance, text analysis

1. Introduction

Understanding the aboutness of Web documents is important for many Web based systems. While in the past the aboutness of documents has been primarily formulated as a problem of finding the top-K most relevant keywords in the document (Tomokiyo and Hurst, 2003; Hulth, 2003) it is a recent trend that many NLP systems are moving towards understanding documents in terms of entities. Currently, there are many named entity recognition systems (NER). Nevertheless, in long texts, the list of recognized entities can be very large containing also entities which are not relevant to the document. Identifying a subset of salient entities that play an important role in a story that the document describes can significantly help to better understand the aboutness of the document. To this end, the research community has explored already existing datasets for the entity salience task, e.g. the Microsoft Document Aboutness (MDA) dataset, and the New York Times (NYT) dataset. However, neither dataset provides the underlying document content due to copyright restrictions. Moreover, in these datasets the entity candidates have been generated automatically using proprietary NER systems.

To the best of our knowledge, in this paper we present the first publicly available corpus with crowdsourced entity salience annotations. The corpus is created by re-using the NEL Reuters–128 corpus (Röder et al., 2014) and published in the NLP Interchange Format (NIF) ensuring high interoperability and re-use. The corpus is available for download under an open license.

To validate and show the potential of the dataset we develop an entity salience learning use case. The use case is based on the crowdsourced dataset and it leverages the local and global information about entities derived from the document itself and the DBpedia knowledge graph. We have conducted several experiments and report the results.

The remainder of this paper is organized as follows. Section 2 motivates our work by reviewing the availability of entity salience corpora. Section 3 describes the methodology we followed to develop our entity salience corpus. Section 4 describes the use case showing how this dataset can be used for training an entity salience classifier. It also describes the experiments we run to validate and evaluate the developed method. Finally, Section 5 concludes the paper and provides a prospect for future enhancements.

2. State of the art

Identification of salient entities is a relatively new research problem and currently there is a lack of resources for learning entity salience. While some datasets have been recently published, in particular the "Microsoft Document Aboutness" (MDA) (Gamon et al., 2013) and the "New York Times" (NYT) (Gillick and Dunietz, 2014), they are either not freely available, manually checked or they do not provide all the necessary content for learning entity salience.

The MDA dataset consists of entities occurring in randomly sampled Web pages (from head and tail distributions) together with a salience assessment for the entities. However, due to copyright restrictions, the underlying textual content is not distributed along with the annotations, so the design of the dataset does not foster its straightforward use.

The NYT dataset is another dataset which has been created as an extension the New York Times dataset. This dataset also, due to copyright restrictions, does not provide the underlying textual content along with the annotations. Moreover, the entity candidates have been generated automatically using a proprietary NER system, under the

assumption that the annotations are 100% correct. Furthermore, the salience annotations in the NYT dataset have also been automatically generated by aligning the entities in the abstract and the document under the assumption that every entity which occurs in the abstract is salient.

3. Crowdsourcing Entity Salience Annotations

In this section, we present the methodology of crowdsourcing the entity salience corpus as well as its specific features.

Our aim is to provide an entity salience corpus that is complete, publicly available, and manually evaluated by humans. To this end, we take up a recently published dataset Reuters-128 (Röder et al., 2014) and crowdsourced the entity salience annotations. Reuters-128 is an English corpus stored in the NIF format, containing 128 economic news articles and has been initially proposed as a dataset for evaluation of NER systems. The dataset provides information for 878 named entities with their position in the document (begin and end offset) and a URI of a DBpedia resource identifying the entity. Since the dataset only provides information about named entities found in the corpus, we have further extended the dataset with "common entities" using Entityclassifier.eu NER system (Dojchinovski and Kliegr, 2013) resulting in additional 3,551 entities. To obtain a gold standard of entity salience judgments we have used the crowdsourcing tool CrowdFlower[1] to collect judgments from non-expert paid judges. The annotators have been given text and a highlighted entity, which has to be classified, following the approach taken by Gamon et al. (2013), into one of the following three classes:

- Most Salient - indicates that that document is mostly about entity, or entity plays a prominent role in the content of the document.
- Less Salient - entity plays an important role in some parts of the document.
- Not Salient - the document is not about the entity.

For each named and common entity in the Reuters-128 dataset, we have collected three judgments from annotators based in 15 different countries, including English-speaking countries, such as United Kingdom, Canada and United States. We have also manually created a set of test questions, which helped us to determine contributor's trust score computed by the CrowdFlower platform. Only judgments from contributors with trust score higher than 70% have been included in the ground-truth. If the trust score of a contributor falls bellow 70%, all his/her judgments were disregarded. Each task consisted of 22 questions that had to be answered in order to get the task completed.

Further, from different contributors we collected three judgments per question. Also, in order to collect judgments from different users, we have limited the maximum number of judgments per contributor to 352.

The crowdsourcing took 36 days, and in total we have collected 18,058 judgments from which 14,528 have been considered as "trusted" and 3,530 as "untrusted" judgments. The interannotator agreement between the annotators was 63.66%. Aggregated result is chosen based on the response with the greatest confidence, where the agreement is weighted by contributor's trust score.

Statistics for the crowdsourced dataset are presented in Table 1.

Corpus	Reuters-128
Documents	128
Entity mentions	4,429
Unique entities	2,024
Entities linked to DBpedia	3,194
Most salient entities	804 (18%)
Less salient entities	1,750 (40%)
Not salient entities	1,875 (42%)

Table 1: Size metrics for the Reuters-128 entity salience dataset.

The dataset has been modelled and published in RDF using NLP Interchange Format (NIF) version 2.0 (Hellmann et al., 2013). This assures easy use and high interoperability of the corpus and its annotations. The complete crowdsourced dataset has been converted in the NIF format and published[2] under the Creative Commons Public Domain CC0 license[3].

4. Use Case: Learning Entity Salience

In order to validate and show the potential of the dataset, we developed a particular use case of the dataset in *learning entity salience*. As a prime source for learning we use the generated dataset, complemented with information about entities derived from the DBpedia knowledge graph. Figure 1 illustrates our methodology of learning salience based on local features–derived from the document itself; and global features–derived from the DBpedia knowledge graph.

The set of local features is derived only from the information available within the document. The main assumption behind this set of features is that the salience is function of the document's structure. In other words, these features should elucidate the way authors i) write and structure articles, and ii) distribute the important entities in the document. The full set of local scope features is summarized in Table 2.

Although the evidence of entity salience can be derived effectively from the document content and its structure, extra information, derived from entity knowledge graph such as DBpedia (Lehmann et al., 2014) can improve the accuracy of the learned entity salience model. To this end, we build a set of features derived from information

[1] http://www.crowdflower.com/

[2] http://ner.vse.cz/datasets/entitysalience-collection/

[3] https://creativecommons.org/publicdomain/

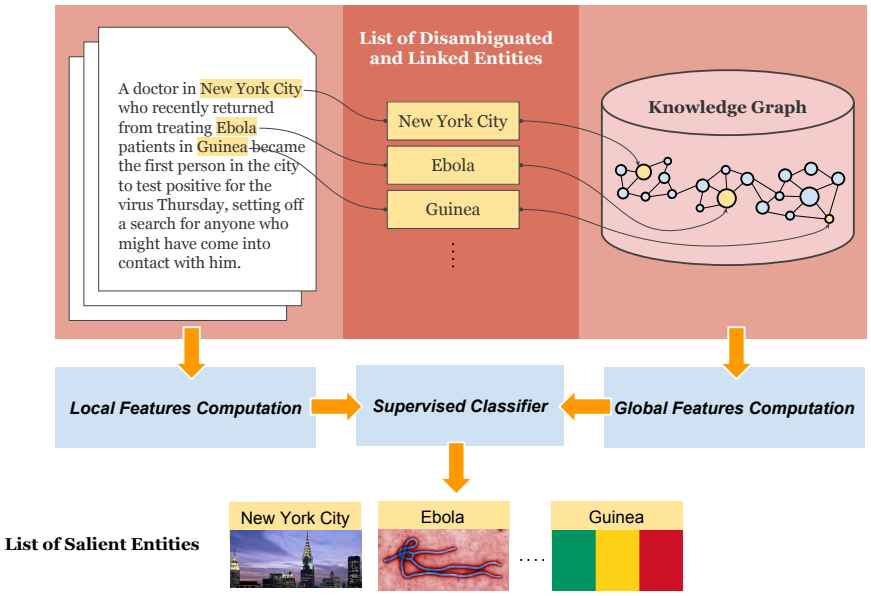

Figure 1: Schematic overview of the methodology for identification of salient entities.

Feature	Type	Description
`entity-type`	nominal	The type of the entity: named entity, common entity.
`1st-begin-index`	numeric	Begin index of the first occurrence of the entity in the document.
`entity-occurrences`	numeric	Number of mentions of the entity in the doc.
`entity-mentions`	numeric	Total number of entity mentions in the document.
`unique-entities`	numeric	Number of unique entities in the document.

Table 2: Features computed from information available within the document.

Feature	Type	Description
`page-rank`	numeric	PageRank value for the given entity.
`hits`	numeric	HITS score for the given entity.
`in-degree`	numeric	Indegree score for the given entity.
`out-degree`	numeric	Outdegree score for the given entity.
`num-triples`	numeric	Number of triples describing the entity.
`num-props`	numeric	Number of properties describing the entity.
`object-props`	numeric	Number of object properties describing the entity.
`datatype-props`	numeric	Number of datatype properties describing the entity.

Table 3: Features computed from DBpedia.

available in DBpedia. There is twofold rationale for inclusion of the global features: first, authors write articles with the assumption that the readers have some background knowledge about the entities; and second, entities popular and central in external knowledge graphs could be also popular and play an important role in a document.

In order to compute the set of global features (see Table 3) we have considered the English DBpedia 2014 pagelinks dataset[4] and the DBpedia knowledge graph. Before computation of the features we have considered cleansing of the DBpedia pagelinks dataset and removed redundant information and links such as redirects which add noise.

Based on this dataset we have computed PageRank (Brin and Page, 1998), Hub and Authorities (HITS) (Kleinberg, 1999), in–degree and out–degree metrics that can be considered as general importance or popularity of an entity. The features with global scope are summarized in Table 3.

[4] `http://data.dws.informatik.uni-mannheim.de/dbpedia/2014/en/page_links_en.ttl.bz2`

4.1. Experiments

We have conducted several experiments in order to validate our dataset and evaluate its potential in learning entity salience. In the experiments we have addressed the following questions:

- *What machine learning algorithm performs the best for our entity salience method?*
- *What is the impact of individual and combined local and global set of features on the performance?*
- *How does the quality of the entity links influence the performance of learning entity salience?*

For the experiments we used the complete Reuters-128 dataset and we performed ten-fold cross validation by partitioning the dataset into ten equal partitions and performing ten cross-validations while training on nine partitions and evaluation on one.

Further, in the experiments we consider the following two baseline methods against which we compare our method.

- *Positional Baseline.* An entity is considered as salient only if the begin index of the first occurrence in the document is within the first 100 characters. This also corresponds to a typical sentence length, which in average is around 100 characters long.
- *Entity Frequency Baseline.* This baseline method is learning from the frequency of the entity in the document. As a learning algorithm, for this method we have used the Random Forest (RF) tree algorithm.

Experiment 1: Comparison of Various Algorithms. In this experiment, we applied various machine learning algorithms to find the most suitable one for learning entity salience. We have experimented with 5 well-known machine learning algorithms: Naive Bayes (NB), Support Vector Machines (SVM) with polynomial kernel, k-Nearest Neighbor (k-NN) with euclidean distance function and k=1, C4.5 decision tree classifier and Random Forest tree classifier with maximum tree depth of 13[5] and the number of trees set to 30.

ML algorithm	Reuters-128		
	Precision	Recall	F1
Naive Bayes	0.518	0.488	0.391
SVM	0.534	0.504	0.416
k-NN	0.566	0.564	0.565
C4.5	0.586	0.586	0.586
Random Forest	**0.612**	**0.608**	**0.607**

Table 4: Results for different learning algorithms.

The results show (see Table 4) that the best performing algorithm is the Random Forest decision tree based classifier with F1 0.607. The second best performance shows the C4.5 decision tree based classifier, with 0.586 F1. The worst performance shows the NaiveBayes classifier with 0.391 F1. For comparison, the Random Forest compared to NaiveBayes shows improvement of nearly 55% which shows that the decision tree based classification algorithms are more suitable for learning entity salience than the simple learning algorithms (k–NN), probabilistic classifiers (NaiveBayes) or associated based learning classifiers (SVM).
The scores reported in Table 4 are computed as "weighted average" for all classes. In Table 5 we report the scores for the "most salient" class only and compares our approach to the baseline methods.

Method	Reuters-128		
	Precision	Recall	F1
Positional baseline	0.518	0.488	0.391
Entity frequency baseline	0.437	0.133	0.204
Our with Random forest	0.693	0.516	0.592

Table 5: Evaluation results for different baseline methods for the "*most salient*" class.

[5]The tree depth 13 corresponds to the number of used features.

Experiment 2: Feature Analysis. In this experiment, we have focused on analysis of the features and their impact on the performance. We have evaluated the contribution of each feature with local and global scope. The contribution of each individual feature has been evaluated by incrementally adding new features, starting with the `begin-index` feature for the local feature set, and `page-rank` for the global feature set. Table 6 summarizes the results from the evaluation.
The results show that the model based on the local features achieves better results than the model based on the global features. It can be also observed that a model which considers both, the local and the global features, achieves better results than model which considers the local or the global features only. Nevertheless, both, the global and the local features achieve sufficiently good performance to be considered individually, when one or another feature family is not available.

Features	Reuters-128		
	Precision	Recall	F1
`1st-begin-index`	0.496	0.493	0.492
`+ entity-occurrences`	0.533	0.531	0.530
`+ entity-mentions`	0.568	0.566	0.565
`+ unique-entities`	0.587	0.585	0.585
`+ entity-type`	0.596	0.593	0.592
`page-rank`	0.451	0.455	0.394
`+ hits`	0.514	0.497	0.461
`+ in-degree`	0.516	0.504	0.483
`+ out-degree`	0.505	0.497	0.481
`+ num-triples`	0.520	0.508	0.492
`+ num-properties`	0.524	0.509	0.492
`+ num-object-properties`	0.522	0.507	0.490
`+ num-datatype-properties`	0.521	0.507	0.489
`all combined`	**0.612**	**0.608**	**0.607**

Table 6: Evaluation results for different features.

Experiment 3: Impact of the Entity Linking. The quality of the links identifying the entities is crucial for computation of different features. Based on these URIs we can compute the number of occurrences of an entity, or count the number of unique entities. Moreover, by linking the entities with an external knowledge graph we can compute various graph metrics, such as the PageRank, HITS or in/out degree.
To evaluate the impact of the quality of entity linking, we randomly created incorrect links in the dataset. Five versions of the dataset were created with different amount of incorrectly linked entities. According to the results from the first experiment, for this experiment, we trained the models using the Random Forest learning algorithm. The results from the experiment are presented in Table 7.

The results from the experiment show that the entity salience learning is influenced by the quality of the entity linking where 10% of incorrectly linked entities results in

Portion of incorrect links	Precision	Recall	F1
`all correct`	0.612	0.608	0.607
`1/10`	0.597	0.592	0.592 (-2.47%)
`1/5`	0.588	0.583	0.582 (-4.12%)
`1/4`	0.587	0.583	0.582 (-4.12%)
`1/3`	0.571	0.566	0.565 (-6.92%)
`1/2`	0.560	0.555	0.553 (-8.90%)

Table 7: Results for different portions of incorrectly linked entities.

2.47% decrease, while 20-25% of incorrectly linked entities result in 4.12% decrease of the accuracy of the salience learning. It can be observed, that even with 50% of incorrectly linked entities, the learning accuracy still shows promising results (F1=0.553). This shows that the features which are not dependent on the entity links (`1st-begin-index`, `entity-mentions`, `entity-type`) can balance the incorrectly linked entities.

5. Conclusion

In the recent years, named entity recognition systems gained great popularity on the Web. However, these systems, at large, do not evaluate the actual importance of the recognized entities in the documents. Since we tackle relatively new research problem, the availability of resources for learning and evaluation of entity salience is very limited. To this end, through crowdsourcing we created the first publicly available entity salience dataset. The dataset is published in the NIF format which fosters interoperability and re-use. We validated the dataset on an entity salience learning use case and reported results from several experiments.

In our future work, we want to extend the feature set with additional information about the entities such as their type information. We also plan to adapt and apply our entity salience model on different types of texts such as microposts, video subtitles and music lyrics.

Acknowledgement. This research was supported by the European Union's 7th Framework Programme via the LinkedTV project (FP7-287911). Tomáš Kliegr was partly supported by the Faculty of Informatics and Statistics, University of Economics, Prague within "long term institutional support for research activities".

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117.

Dojchinovski, M. and Kliegr, T. (2013). Entityclassifier.eu: Real-time classification of entities in text with wikipedia. In Hendrik Blockeel, et al., editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8190 of *Lecture Notes in Computer Science*, pages 654–658. Springer Berlin Heidelberg.

Gamon, M., Yano, T., Song, X., Apacible, J., and Pantel, P. (2013). Identifying salient entities in web pages. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2375–2380. ACM.

Gillick, D. and Dunietz, J. (2014). A new entity salience task with millions of training examples. In *Proceedings of the European Association for Computational Linguistics*.

Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. (2013). Integrating nlp using linked data. In Harith Alani, et al., editors, *The Semantic Web ÃćÂĂÂŞ ISWC 2013*, volume 8219 of *Lecture Notes in Computer Science*, pages 98–113. Springer Berlin Heidelberg.

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 216–223, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, September.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., et al. (2014). DBpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*.

Röder, M., Usbeck, R., Hellmann, S., Gerber, D., and Both, A. (2014). N3 - a collection of datasets for named entity recognition and disambiguation in the nlp interchange format.

Tomokiyo, T. and Hurst, M. (2003). A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18*, MWE '03, pages 33–40, Stroudsburg, PA, USA. Association for Computational Linguistics.