Document Aboutness via Entity Annotation

Marco Ponza, Paolo Ferragina, and Francesco Piccinno

Dipertimento di Informatica, University of Pisa
{lastname}@di.unipi.it

Abstract. very very short.

Keywords: entity salience, document aboutness, information extraction

1 Introduction

In Information Retrieval (IR), *document aboutness* is the problem that asks for creating a succinct representation of a document's subject matter [?,?,?] via keywords, named entities, concepts, snippets or sentences. Current solutions are mainly based on two algorithmic approaches: the first and classic one, known as keyphrase extraction [?,?], represents the abountness of an input document by its *lexical elements* such as keywords, terms or sentences. The second and more recent one, known as *entity salience* [?], captures the aboutness of an input document by its *semantic elements* represented via entities drawn from a Knowledge Base, such as Freebase or Wikipedia. This latter approach strongly depends on the effectiveness and efficiency of the *entity annotation* process, which currently offers several performant solutions [?,?,?,?] and easily overcomes the limitations of the approaches based on keyphrase extraction, which will be discussed in Section **??**.

The state-of-the-art in entity salience is offered by the CMU-Google's system [?] which achieves precision and recall of 60.5% and 63.5%, respectively, over the New York Time dataset.[1] This system uses a proprietary entity annotator to extract entities from documents and a very simple binary classifier to distinguish between salient and non-salient entities. More specifically, each entity is mapped into a feature space which depends on the position of its first mention, frequency of the mentions (using a coreference resolver), lower-cased head word, part-of-speech tags and centrality of entities in the Freebase's graph. Authors show that their model significantly outperforms a baseline model based just on sentence position but conclude the paper by stating that: There is likely significant room for improvement, especially by leveraging background information about the entities, [...]. Perhaps features more directly linked to Wikipedia, as in related work on keyword extraction, can provide more focused background information."

[1] PAOLO: Perch non F1? Sotto si usa micro-F1, altrimenti il lettore non sa fare un confronto.

In our paper we address these issues by designing and implementing a new system for entity-salience detection, called Swat, that is based on the careful orchestration of sophisticated and state-of-the-art tools publicily available in the IR and NLP literature that will allow to extract several new features from both the lexical and the semantic elements of the input document. In addition, we aim to design a system which is publicily available (unlike [?] which deploys proprietary modules), may work on arbitrary types of texts (unlike [?] which is tailored to news), and improves the CMU-Google's system over the large New York Times (NYT) dataset [?].

Specifically, the algorithmic structure of Swat will be based on three building blocks: (1) CoreNLP, the most well-known NLP framework to analyze the grammatical structure of sentences (e.g., it is able to detect subject or object of verbs); (2) Wat [?], one of the best publicily available entity annotators [?] which implements the whole annotation pipeline (namely mention detection and entity linking, the so called A2W task); (3) TextRank [?], the popular document summarizer which ranks sentences by a relevance score that is computed running PageRank over a graph built upon the structure of the input document. These three tools will be orchestrated to extract semantic and lexical information from the input documents so to derive a large and sophisticate set of novel features upon which we will design and implement a binary classifier able to efficaciously select only the salient entities.

The performance of our system Swat will be checked via a large experimental test over the well-known annotated New York Times (NYT) dataset [?], consisting of 110,540 news for a total 1.4 GB of textual data, where each document contains an average of 975 tokens per document. BY a careful study of the effectiveness of each proposed feature and

The final result will show that Swat raises the best known performance of state-of-the-art systems to 62.6% in terms of micro-F1. This is an absolute improvement of 0.6%, which becomes very significant (and robust) given the size of the NYT dataset and the fact that Swat detects 7.000 more salient entities over an overall detection of 120.000 salient ones (+6%).[2]

Overall, the main contributions of the paper are the following ones:[3]

- We design Swat, a novel and effective entity-salience system that deploys a rich and novel set of lexical and semantic features to identify the semantic focus of a document. In order to encourage and support reproducible scientific research on this task, we release Swat as a public available tool.
- The design of Swat required a throughtful study over a novel and rich set of features capturing the semantic (i.e. entities of Wikipedia) and structured

[2] PAOLO: Questo molto delicato lo perfezioniamo quando tutti i risultati sono chiari. Ci giochiamo molto in questo passaggio. Il miglioramento pu essere poco ma occorre evidenziare che il sistema disponibile via API e poi trovare altro... Il miglioramento sull'insieme bilanciato?

[3] PAOLO: occorre rendere tutto WOW, nel senso che ogni cosa deve dire per bene dove ci differenziamo con Google.

(i.e. dependency trees) content of the input document, as hoped for in the conclusions of [?].

- Swat improves the state-of-the-art CMU-Google's system [?] both in terms of micro- and macro-F1 over the large and well-known New York Times (NYT) dataset [?]. The improvement is thus robust and is accompayned by a throughtful analysis of the novel features and an investigation of the erroneous predictions of Swat which will finally allow us to identify pro/cons of the proposed approach which deserve attention in the future.

2 Related Work

In recent years, the automatic identification of *aboutness* in documents has been addressed by several kinds of algorithmic approaches, differing both on its representation (e.g. keyphrases, named entities or Wikipedia pages) and on the algorithms deploied. Nevertheless, all these approaches usually work into two main phases: The first one aims at generating a list of candidates (whose nature depends on the chosen representation); the second one aims at ranking or classifying those candidates by some relevance scores properly computed from lexical and/or semantic features extracted from the input document.

Classical aboutness approaches are based on *keyphrase extraction* algorithms which extract salient phrases from the input documents (aka keyphrases). In this context, the generation of candidate keyphrases is usually performed by applying different heuristics [?], such as: considering as candidates only words labeled with specific POS-tags [?,?,?], extracting n-grams [?,?], keeping only those phrases which belong to a fixed dictionary of terms [?] or by considering only word sequences which have been identified as proper nouns [?]. In a second phase thos systems select the salient keyphrases from the candidate ones via supervised or unsupervised approaches [?,?,?]. The former ones work on a feature space mainly composed by statistical (e.g., frequency, inverse document frequency, etc.) and positional (e.g. position of the first/last keyphrase mention, etc.) features. The latter ones build a graph where nodes are candidate keyphrases and edges are weighted by their co-occurence frequency in the input document. The salience scores are then computed by running graph-based centrality measures [?,?,?].

Unfortunately, keyphrase extraction systems show several obvious limitations [?]. Firstly, the aboutness expressed by keyphrase extractors is ambiguous: the sequence of words leave their interpretation to the reader! For example, if the extracted keyphrase is *Michael Jordan*, we cannot be sure whether the document is about the basket player or the professor. A second limitation consists of the fact that the extracted keyphrases are far to be perfect [?,?]: words that appear frequently in the input document may possibly induce the selection of not-salient phrases (aka, *overgeneration errors*); infrequent keyphrases go undetected (aka, *infrequency errors*); working at a pure lexical level those systems are not able to detect the semantic equivalence between two keyphrases (aka, *redundancy errors*).

The modern trend in entity annotators [?,?,?,?] has been exploited very recently by some researchers [?,?] in order to introduce some "semantic" into the document aboutness representation and thus to overcome the limitations above. The key idea is to represent the aboutness of the input document via well-defined concepts such as the entities which occur in that document and are drawn from a knowledge base (e.g. Wikipedia, Wikidata, Freebase). This way the entities are unambiguous and, moreover, can exploit the structure of the graph underling those KBs in order to implement new extraction algorithms or empower their document representation for kinds of applications ranging from document clustering or classification [?,?] to exploratory searches [?], up to Contextual ads or document analysis [?,?], just to cite a few.

The state of the art in entity salience is [?]. Our work differs from this one in several aspects, some of which have been sketched in the previous section: our system SWAT is designed over three novel modules (TEXTRANK, WAT and CORENLP) which allow us to extract a variegate and new powerful set of lexical and semantic features, these features are not tailored to news documents and so SWAT could be applied to any kinds of texts, we perform a throughtful and robust analysis about the impact of those features in SWAT's performance and over the large NYT's dataset. Finally we remark that, unlike [?], our system SWAT is designed over SW components which are publicily available so it will be released to the public as an open-source project.

Our work also differs from [?] in several aspects. In terms of features, we experience with new lexical features such as the powerful TextRank and Dependency Trees, whereas for the semantic features we experience with many other centrality measures over the graph of entities such as Katz, HITS, Harmonic, etc.; this clearly provides a wider analysis of the document aboutness problem which fully addresses the hope expressed in the conclusions of [?], as we mentioned in the previous section. In terms of algorithmic structure, the approach in [?] is supervised both in the annotation and in the salience detection step, furthermore it solves the salience taking into account the lexical content of the spots pointing to the entities ("it is possible to have spots without any predicted relevant entity, and spots with more than one relevant entity"); conversely we solve the salience over the spot-entity annotation thus being able to distinguish the salience of two different entities which are mentioned by the same spot (i.e. New York e New York Yankee). In terms of experiments [?] compares its performance against *pure* entity annotators (such as TagMe), and not the state-of-the-art system in entity salience of [?], and moreover they do not use the publicily available NYT's dataset but confine the experiments over 365 news versus the more than 110,000 news test of [?]. Our system SWAT will be compared against the state-of-the-art [?] over the large NYT's dataset.