A New Entity Salience Task with Millions of Training Examples

Jesse Dunietz
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213, USA
jdunietz@cs.cmu.edu

Dan Gillick
Google Research
1600 Amphitheatre Parkway
Mountain View, CA 94043, USA
dgillick@google.com

Abstract

Although many NLP systems are moving toward entity-based processing, most still identify important phrases using classical keyword-based approaches. To bridge this gap, we introduce the task of *entity salience*: assigning a relevance score to each entity in a document. We demonstrate how a labeled corpus for the task can be automatically generated from a corpus of documents and accompanying abstracts. We then show how a classifier with features derived from a standard NLP pipeline outperforms a strong baseline by 34%. Finally, we outline initial experiments on further improving accuracy by leveraging background knowledge about the relationships between entities.

1 Introduction

Information retrieval, summarization, and online advertising rely on identifying the most important words and phrases in web documents. While traditional techniques treat documents as collections of keywords, many NLP systems are shifting toward understanding documents in terms of entities. Accordingly, we need new algorithms to determine the prominence – the *salience* – of each entity in the document.

Toward this end, we describe three primary contributions. First, we show how a labeled corpus for this task can be automatically constructed from a corpus of documents with accompanying abstracts. We also demonstrate the validity of the corpus with a manual annotation study. Second, we train an entity salience model using features derived from a coreference resolution system. This model significantly outperforms a baseline model based on sentence position. Third, we suggest how our model can be improved by leveraging background information about the entities and their relationships – information not specifically provided in the document in question.

Our notion of salience is similar to that of Boguraev and Kenney (1997): “discourse objects with high salience are the focus of attention”, inspired by earlier work on Centering Theory (Walker et al., 1998). Here we take a more empirical approach: salient entities are those that human readers deem most relevant to the document.

The entity salience task in particular is briefly alluded to by Cornolti et al. (2013), and addressed in the context of Twitter messages by Meij et. al (2012). It is also similar in spirit to the much more common keyword extraction task (Tomokiyo and Hurst, 2003; Hulth, 2003).

2 Generating an entity salience corpus

Rather than manually annotating a corpus, we automatically generate salience labels for an existing corpus of document/abstract pairs. We derive the labels using the assumption that the salient entities will be mentioned in the abstract, so we identify and align the entities in each text.

Given a document and abstract, we run a standard NLP pipeline on both. This includes a POS tagger and dependency parser, comparable in accuracy to the current Stanford dependency parser (Klein and Manning, 2003); an NP extractor that uses POS tags and dependency edges to identify a set of entity mentions; a coreference resolver, comparable to that of Haghighi and Klein, (2009) for clustering mentions; and an entity resolver that links entities to Freebase profiles. The entity resolver is described in detail by Lao, et al. (2012).

We then apply a simple heuristic to align the entities in the abstract and document: Let M_E be the set of mentions of an entity E that are proper names. An entity E_A from the abstract aligns to an entity E_D from the document if the syntactic head token of some mention in M_{E_A} matches the head token of some mention in M_{E_D}. If E_A aligns with more than one document entity, we align it with the document entity that appears earliest.

In general, aligning an abstract to its source document is difficult (Daumé III and Marcu, 2005).

We avoid most of this complexity by aligning only entities with at least one proper-name mention, for which there is little ambiguity. Generic mentions like *CEO* or *state* are often more ambiguous, so resolving them would be closer to the difficult problem of word sense disambiguation.

Once we have entity alignments, we assume that a document entity is salient only if it has been aligned to some abstract entity. Ideally, we would like to induce a salience ranking over entities. Given the limitations of short abstracts, however, we settle for binary classification, which still captures enough salience information to be useful.

2.1 The New York Times corpus

Our corpus of document/abstract pairs is the annotated New York Times corpus (Sandhaus, 2008). It includes 1.8 million articles published between January 1987 and June 2007; some 650,000 include a summary written by one of the newspaper's library scientists. We selected a subset of the summarized articles from 2003-2007 by filtering out articles and summaries that were very short or very long, as well as several special article types (e.g., corrections and letters to the editor).

Our full labeled dataset includes 110,639 documents with 2,229,728 labeled entities; about 14% are marked as salient. For comparison, the average summary is about 6% of the length (in tokens) of the associated article. We use the 9,719 documents from 2007 as test data and the rest as training.

2.2 Validating salience via manual evaluation

To validate our alignment method for inferring entity salience, we conducted a manual evaluation. Two expert linguists discussed the task and generated a rubric, giving them a chance to calibrate their scores. They then independently annotated all detected entities in 50 random documents from our corpus (a total of 744 entities), without reading the accompanying abstracts. Each entity was assigned a salience score in $\{1, 2, 3, 4\}$, where 1 is most salient. We then thresholded the annotators' scores as salient/non-salient for comparison to the binary NYT labels.

Table 1 summarizes the agreement results, measured by Cohen's kappa. The experts' agreement is probably best described as *moderate*,[1] indicating that this is a difficult, subjective task, though deciding on the most salient entities (with score 1) is easier. Even without calibrating to the induced NYT salience scores, the expert vs. NYT agreement is close enough to the inter-expert agreement to convince us that our induced labels are a reasonable if somewhat noisy proxy for the experts' definition of salience.

Comparison	$\kappa_{\{1,2\}}$	$\kappa_{\{1\}}$
A1 vs. A2	0.56	0.69
A1 vs. NYT	0.36	0.48
A2 vs. NYT	0.39	0.35
A1 & A2 vs. NYT	0.43	0.38

Table 1: Annotator agreement for entity salience as a binary classification. A1 and A2 are expert annotators; NYT represents the induced labels. The first κ column assumes annotator scores $\{1, 2\}$ are salient and $\{3, 4\}$ are non-salient, while the second κ column assumes only scores of 1 are salient.

3 Salience classification

We built a regularized binary logistic regression model to predict the probability that an entity is salient. To simplify feature selection and to add some further regularization, we used feature hashing (Ganchev and Dredze, 2008) to randomly map each feature string to an integer in $[1, 100000]$; larger alphabet sizes yielded no improvement. The model was trained with L-BGFS.

3.1 Positional baseline

For news documents, it is well known that sentence position is a very strong indicator for relevance. Thus, our baseline is a system that identifies an entity as salient if it is mentioned in the first sentence of the document. (Including the next few sentences did not significantly change the score.)

3.2 Model features

Table 2 describes our feature classes; each individual feature in the model is a binary indicator. Count features are bucketed by applying the function $f(x) = \text{round}(\log(k(x + 1)))$, where k can be used to control the number of buckets. We simply set $k = 10$ in all cases.

3.3 Experimental results

Table 3 shows experimental results on our test set. Each experiment uses a classification threshold of 0.3 to determine salience, which in each case is very close to the threshold that maximizes F_1. For comparison, a classifier that always predicts the majority class, non-salient, has $F_1 = 23.9$ (for the *salient* class).

[1] For comparison, word sense disambiguation tasks have reported agreement as low as $\kappa = 0.3$ (Yong and Foo, 1999).

Feature name	Description
`1st-loc`	Index of the sentence in which the first mention of the entity appears.
`head-count`	Number of times the head word of the entity's first mention appears.
`mentions`	Conjuction of the numbers of named (*Barack Obama*), nominal (*president*), pronominal (*he*), and total mentions of the entity.
`headline`	POS tag of each word that appears in at least one mention and also in the headline.
`head-lex`	Lowercased head word of the first mention.

Table 2: The feature classes used by the classifier.

Lines 2 and 3 serve as a comparison between traditional keyword counts and the mention counts derived from our coreference resolution system. Named, nominal, and pronominal mention counts clearly add significant information despite coreference errors. Lines 4-8 show results when our model features are incrementally added. Each feature raises accuracy, and together our simple set of features improves on the baseline by 34%.

4 Entity centrality

All the features described above use only information available within the document. But articles are written with the assumption that the reader knows something about at least some of the entities involved. Inspired by results using Wikipedia to improve keyword extraction tasks (Mihalcea and Csomai, 2007; Xu et al., 2010), we experimented with a simple method for including background knowledge about each entity: an adaptation of PageRank (Page et al., 1999) to a graph of connected entities, in the spirit of Erkan and Radev's work (2004) on summarization.

Consider, for example, an article about a recent congressional budget debate. Although House Speaker John Boehner may be mentioned just once, we know he is likely salient because he is closely related to other entities in the article, such as Congress, the Republican Party, and Barack Obama. On the other hand, the Federal Emergency Management Agency may be mentioned repeatedly because it happened to host a major presidential speech, but it is less related to the story's key figures and less central to the article's point.

#	Description	P	R	F_1
1	Positional baseline	59.5	37.8	46.2
2	`head-count`	37.3	54.7	44.4
3	`mentions`	57.2	51.3	54.1
4	`1st-loc`	46.1	60.2	52.2
5	+ `head-count`	52.6	63.4	57.5
6	+ `mentions`	59.3	61.3	60.3
7	+ `headline`	59.1	61.9	60.5
8	+ `head-lex`	59.7	63.6	61.6
9	+ `centrality`	60.5	63.5	62.0

Table 3: Test set (P)recision, (R)ecall, and (F) measure of the *salient* class for some combinations of features listed in Table 2. The `centrality` feature is discussed in Section 4.

Our intuition about these relationships, mostly not explicit in the document, can be formalized in a local PageRank computation on the entity graph.

4.1 PageRank for computing centrality

In the weighted version of the PageRank algorithm (Xing and Ghorbani, 2004), a web link is considered a weighted vote by the containing page for the landing page – a directed edge in a graph where each node is a webpage. In place of the web graph, we consider the graph of Freebase entities that appear in the document. The nodes are the entities, and a directed edge from E_1 to E_2 represents $P(E_2|E_1)$, the probability of observing E_2 in a document given that we have observed E_1. We estimate $P(E_2|E_1)$ by counting the number of training documents in which E_1 and E_2 co-occur and normalizing by the number of training documents in which E_1 occurs.

The nodes' initial PageRank values act as a prior, where the uniform distribution, used in the classic PageRank algorithm, indicates a lack of prior knowledge. Since we have some prior signal about salience, we initialize the node values to the normalized mention counts of the entities in the document. We use a damping factor d, allowing random jumps between nodes with probability $1 - d$, with the standard value $d = 0.85$.

We implemented the iterative version of weighted PageRank, which tends to converge in under 10 iterations. The `centrality` features in Table 3 are indicators for the rank orders of the converged entity scores. The improvement from adding centrality features is small but statistically significant at $p \leq 0.001$.

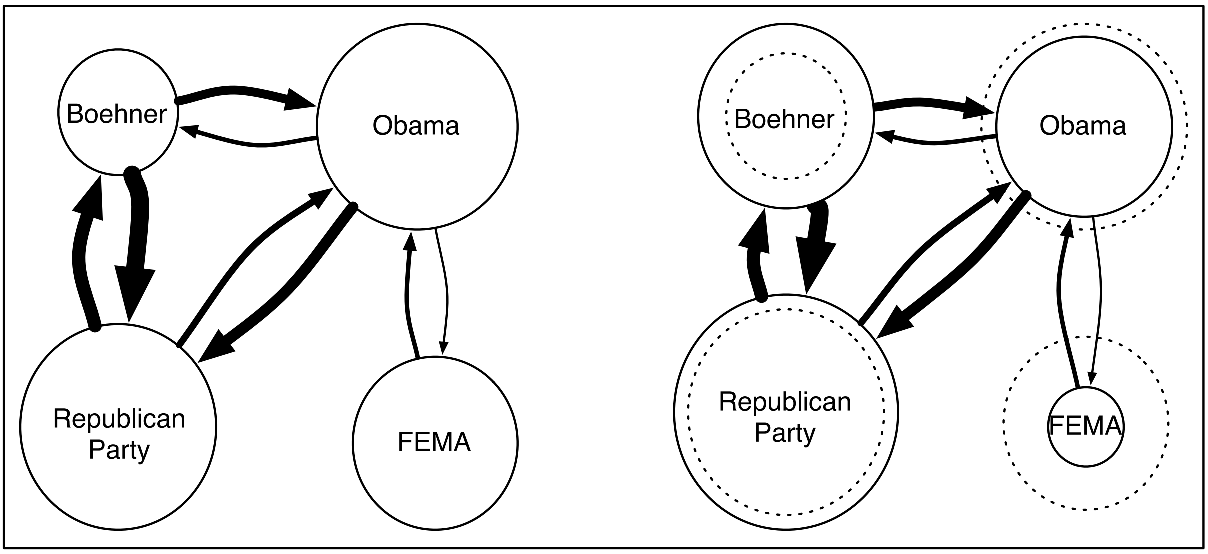

Figure 1: A graphical representation of the centrality computation on a toy example. Circle size and arrow thickness represent node value and edge weight, respectively. The initial node values, based on mention count, are shown on the left. The final node values are on the right; dotted circles show the initial sizes for comparison. Edge weights remain constant.

4.2 Discussion

We experimented with a number of variations on this algorithm, but none gave much meaningful improvement. In particular, we tried to include the neighbors of all entities to increase the size of the graph, with the values of neighbor entities not in the document initialized to some small value k. We set a minimum co-occurrence count for an edge to be included, varying it from 1 to 100 (where 1 results in very large graphs). We also tried using Freebase relations between entities (rather than raw co-occurrence counts) to determine the set of neighbors. Finally, we experimented with undirected graphs using unnormalized co-occurrence counts.

While the ranked centrality scores look reasonable for most documents, the addition of these features does not produce a substantial improvement. One potential problem is our reliance on the entity resolver. Because the PageRank computation links all of a document's entities, a single resolver error can significantly alter all the centrality scores. Perhaps more importantly, the resolver is incomplete: many tail entities are not included in Freebase.

Still, it seems likely that even with perfect resolution, entity centrality would not significantly improve the accuracy of our model. The `mentions` features are sufficiently powerful that entity centrality seems to add little information to the model beyond what these features already provide.

5 Conclusions

We have demonstrated how a simple alignment of entities in documents with entities in their accompanying abstracts provides salience labels that roughly agree with manual salience annotations. This allows us to create a large corpus – over 100,000 labeled documents with over 2 million labeled entities – that we use to train a classifier for predicting entity salience.

Our experiments show that features derived from a coreference system are more robust than simple word count features typical of a keyword extraction system. These features combine nicely with positional features (and a few others) to give a large improvement over a first-sentence baseline.

There is likely significant room for improvement, especially by leveraging background information about the entities, and we have presented some initial experiments in that direction. Perhaps features more directly linked to Wikipedia, as in related work on keyword extraction, can provide more focused background information.

We believe entity salience is an important task with many applications. To facilitate further research, our automatically generated salience annotations, along with resolved entity ids, for the subset of the NYT corpus discussed in this paper are available here:
https://code.google.com/p/nyt-salience/

References

Branimir Boguraev and Christopher Kennedy. 1997. Salience-based content characterisation of text documents. In *Proceedings of the ACL*, volume 97, pages 2–9.

Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. 2013. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd international conference on World Wide Web*, pages 249–260.

Hal Daumé III and Daniel Marcu. 2005. Induction of word and phrase alignments for automatic document summarization. *Computational Linguistics*, 31(4):505–530.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*, 22(1):457–479.

Kuzman Ganchev and Mark Dredze. 2008. Small statistical models by random feature mixing. In *Proceedings of the ACL08 HLT Workshop on Mobile Language Processing*, pages 19–20.

Aria Haghighi and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1152–1161. Association for Computational Linguistics.

Anette Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 216–223.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430.

Ni Lao, Amarnag Subramanya, Fernando Pereira, and William W Cohen. 2012. Reading the web with learned syntactic-semantic inference rules. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1017–1026. Association for Computational Linguistics.

Edgar Meij, Wouter Weerkamp, and Maarten de Rijke. 2012. Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 563–572. ACM.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab.

Evan Sandhaus. 2008. The New York Times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Takashi Tomokiyo and Matthew Hurst. 2003. A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 33–40.

Marilyn A Walker, Aravind Krishna Joshi, and Ellen Friedman Prince. 1998. *Centering theory in discourse*. Oxford University Press.

Wenpu Xing and Ali Ghorbani. 2004. Weighted pagerank algorithm. In *Communication Networks and Services Research*, pages 305–314. IEEE.

Songhua Xu, Shaohui Yang, and Francis Chi-Moon Lau. 2010. Keyword extraction and headline generation using novel word features. In *AAAI*.

Chung Yong and Shou King Foo. 1999. A case study on inter-annotator agreement for word sense disambiguation.