Facts That Matter

Marco Ponza[§], **Luciano Del Corro**[†], **Gerhard Weikum**[†]
[§]Department of Computer Science, University of Pisa, Italy
[†]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
marco.ponza@di.unipi.it, {corrogg,weikum}@mpi-inf.mpg.de

Abstract

This work introduces fact salience: The task of generating a machine-readable representation of the most prominent information in a text document as a set of facts. We also present SALIE, the first fact salience system. SALIE is unsupervised and knowledge agnostic, based on open information extraction to detect facts in natural language text, PageRank to determine their relevance, and clustering to promote diversity. We compare SALIE with several baselines (including positional, standard for saliency tasks), and in an extrinsic evaluation, with state-of-the-art automatic text summarizers. SALIE outperforms baselines and text summarizers showing that facts are an effective way to compress information.

1 Introduction

Automatic knowledge acquisition at large scale requires the transformation of human-readable knowledge into a machine-understandable format. Machine-readable information is usually structured in the form of facts, in which a given relation links a set of arguments [e.g., *("US", "withdraws from", "Iran nuclear deal")*]. Facts are at the core of several natural language understanding applications such as knowledge-base (KB) construction (Nguyen et al., 2017), question answering (Abujabal et al., 2018), structured search (Bast et al., 2014), or entity-linking (Cheng and Roth, 2013).

Different approaches aim to discover facts from natural language text. In the extremes of the spectrum, relation extraction (Mintz et al., 2009) looks for all facts linkable to a KB, whereas open information extraction (Banko et al., 2007) extracts facts over an unconstrained set of arguments and relations. In this paper, we aim to additionally score facts according to their prominence.

We define *fact salience* as the task of discovering the most prominent facts in a text document. A fact is salient if it carries the essential information that the text conveys. A higher salient score denotes higher prominence, determining a ranking across all facts in the document. This ranking must reflect relevance and diversity: We want the top-k facts to compress the most relevant information in the smallest number of facts.

Fact salience is closely related to automatic text summarization (Erkan and Radev, 2004) as both try to capture the essential information in a document. However, fact salience output is required to be interpreted by machines to a certain extent. Text summarization, on the contrary, is meant to be understood by humans alone; it is often composed by ungrammatical text and isolated keywords, making it difficult to structure in a machine-readable form ex-post.

Here we present SALIE (Salient Information Extraction), the first fact salience system able to output a ranking of salient open facts from a text document. SALIE is unsupervised and knowledge agnostic. It uses facts as atomic units and PageRank to detect their relevance. It also exploits the fact structure to promote diversity via clustering.

We evaluated SALIE on a real-world dataset and compared it with the strong positional baseline (facts appearing first are more relevant) and, in an extrinsic evaluation, with two top text summarizers (one reimplemented to work at fact level). SALIE outperforms baselines and text summarization competitors particularly when the size of the output is restricted, suggesting that facts, as atomic units expressing a single proposition (Del Corro and Gemulla, 2013), are an effective way to compress information.

The source code and the processed datasets are

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1043–1048
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

publicly available[1] to encourage further developments of the fact salience task.

2 Fact Salience

Fact salience is the task of extracting *salient* facts from a text document. *Salient* facts must fulfil two requirements: (i) *relevance* and (ii) *diversity*.

A fact is relevant if it carries the essential information that the text conveys. A fact is not relevant per se but in a specific context. In an article about the US-Iran nuclear deal the fact *("US", "withdraws from", "Iran Nuclear Deal")* is more relevant that *("Washington", "is", "US capital")*.

The output of a fact salience system must ensure that the top-k facts contain the maximum information in the smallest number of facts. This implies a dependency between facts as less relevant facts should be penalized when they carry information already contained in more relevant ones.

3 Related Work

Fact salience is close to automatic text summarization (Erkan and Radev, 2004); both must detect the most prominent information in the text. However, while text summarization generates summaries for humans, fact salience output must be interpretable by machines. Fluency and language cohesion are not requirements for fact salience.

Triple scoring in KBs (Bast et al., 2017) is also related. However, while in fact salience a fact is not relevant per se but locally in a textual context, triple KB scoring asses the global relevance of a KB fact for a specific entity [*("T. Burton", "profession", "actor")* vs. *("T. Burton", "profession", "director")*] .

Typically, a text summarization system splits the text into atomic units (usually sentences) that are scored and ranked (Allahyari et al., 2017). Diversity is generally guaranteed by clustering them in topics and selecting the most representative members from each cluster. Once selected, the atomic units are compressed to ensure minimality.

Generating a machine-readable representation from text summarization output is difficult. This output can be incomplete or ungrammatical, given the use of compression techniques (Zajic et al., 2007), or the inclusion of keywords or short unconnected phrases with topical information (Hasan and Ng, 2014). Open information extractors will most likely fail to generate meaningful facts in these circumstances. However, text summarization techniques to score the atomic units can be exploited for fact salience.

Open facts have been already used in text summarization for redundancy, using synonymity (Christensen et al., 2013) or as input for a classifier (Christensen et al., 2014). In this case, we use facts as atomic units. Working at the fact level provides a natural framework to detect essential information in a text document, since facts are minimal comprehensive atomic units expressing a single proposition (Del Corro and Gemulla, 2013). This helps to avoid working with sentences that might express more than one proposition or arbitrary chunking the input text. Compression is also more principled at a fact level as the fact hierarchical structure is clearly defined (Gashteovski et al., 2017). Additionally, we exploit the fact structure to promote diversity.

Several supervised and unsupervised methods have been used in text summarization to determine the relative prominence of the atomic units. For instance, two of the top performer systems Durrett et al. (2016) and Mihalcea and Tarau (2004), which we include in our extrinsic evaluation, are based on ILP and an unsupervised graph algorithm respectively. Other approaches include LDA (Pouriyeh et al., 2017), ontology-based (Baralis et al., 2013) or clustering (Yang et al., 2014), and more recently neural-based methods (See et al., 2017). As in Mihalcea and Tarau (2004) or Erkan and Radev (2004) we use PageRank to establish the relative prominence of the atomic units (Sec. 4.1). However, we weight the graph edges using word vectors to allow more expressive semantics, avoiding the sparsity of frequency-based methods.

Different approaches have also been explored to promote diversity. Xiong and Luo (2014), for example, use LSA, and Chien and Chang (2013) rely on topic models. In our case, we generate diversity by exploiting the fact structure (Sec. 4.2). We cluster facts in terms of their subjects as a way to have the most relevant information about the different entities appearing in the text. The subject is typically the topic of the clause or proposition (Quirk et al., 1985).

[1] https://github.com/mponza/SalIE

4 SALIE: Salient Information Extraction

SALIE is a graph-based method for the extraction of *salient* open facts in text documents. Open facts are a structured machine-readable representation of the information in text. Its arguments are not linked to an existing KB. SALIE takes as input all open facts detected by an open information extraction system (in our implementation we use MINIE (Gashteovski et al., 2017)).

SALIE works in two stages: (i) relevance and (ii) diversification. First, a graph with open facts as nodes is instantiated so that PageRank assesses their relative relevance. Later, a clustering algorithm selects a diversified set of facts.

4.1 Fact Relevance

SALIE computes fact *relevance* by growing a complete graph of open facts $G_{OF} = (V, E)$ extracted from the input text. *Coherence* is induced by weighting the edges E between nodes V, whereas a *relevance* prior is induced via the instantiation of the PageRank's teleport vector.

Step 1 – Facts as Nodes. Each node is a fact extracted by MINIE. Undefined facts (with no clear co-reference) [*("He", "plays", "softball")*] or facts with constituents composed by single words (generally uninformative or noisy) [*("doorman", "has", "age")*] are removed.

Step 2 – Coherence: Edge Weighting. We want related facts to get a higher weight assuming that the most relevant facts will be those more central. We weight each edge (u, v) with the *semantic similarity* between u and v as the cosine between the centroid of the word embeddings in the facts. Stanovsky et al. (2015) have shown that learning word embeddings with open facts allows the generation of higher quality vectors . The assumption is that the relatedness of words within a fact is stronger than with words outside. This provides the basis for more accurate contextualization. Accordingly, in our implementation we use GloVe (Pennington et al., 2014) trained on the Wikipedia corpus using open facts extracted by MINIE for co-occurence context.

Step 3 – Relevance Prior. We introduce a prior for each fact by computing a score used to instantiate the PageRank's teleport vector. The assumption is that authors tend to express the most relevant facts at the beginning. We instantiated each fact teleport as $factPrior(i) = \frac{x_i}{\|X\|}$, where $x_i = |V| - i$ and i is the fact index. This is important especially for news where the lead paragraph is the most important part of the article. That's why the positional baseline is so strong in tasks as text summarization or entity salience (Ponza et al., 2018).

Step 4 – Relevance Computation. This stage runs PageRank on the graph. The stationary distribution will capture the *relevance* of each open fact. This distribution reflects the semantic centrality of each fact weighted by its relevance prior.

4.2 Fact Diversification

In this stage SALIE diversifies the set of *relevant* facts computed in the previous stage. Facts are clustered exploiting the fact structure, and the most relevant facts in each cluster are selected according to the relevance scores.

Facts have clear semantics regarding the role of each of its constituents (i.e., subject, relation, and object) in the proposition. SALIE exploits this by clustering together those facts that have the same head in the subject's constituent. As the subject is typically the theme (or topic) of a the clause (Quirk et al., 1985), the intuition here is that facts with the same subject express information about the same entity. Therefore, each cluster will contain a ranked set of facts about each entity in the document.

After the facts have been clustered, we iteratively select facts from each cluster according to its relevance until we reach the desired number of facts as output. The number of facts in the output is a parameter of the system.

5 Experiments

Methodology. Given a document we want to evaluate how salient the top-k facts are. The number of facts in the ranking is a parameter of the model so we evaluate 5 configurations: top-1 to top-5 facts.
Dataset. As there is no dataset to directly asses the saliency of facts, we compare the extracted facts in each ranking with a manually generated summary. We use the New York Times (Sandhaus, 2008) corpus, consisting of 3956 news articles and summaries from 2007 (with summaries larger than 50 tokens) as described by Durrett et al. (2016).
Metrics. To measure how close is the ranking to the summary, we use the ROUGE package[2], standard for document summarization (Lin,

[2] pypi.org/project/pyrouge/0.1.3

Human Summary		
Body of Toni Grossi Abrams, widow and Staten Island socialite, is found in warehouse on outskirts of Panama City, Panama, where she had moved to begin career in real estate; Debra Ann Ridgley, one of her tenants, is charged with stabbing Abrams to death in her apartment on April 9.		
Method		**Salient Facts / Summary**
Position	1	*("Surgery patients", "lie low in", "style retreat")*
	2	*("Remains", "were discovered beside warehouse at", "edge of cinder-topped soccer field on outskirts of Panama City")*
	3	*("Abrams", "had been stabbed to death in", "apartment")*
TextRank	1	*("Ridgley", "was in Abrams's apartment", "Garcia and friend")*
	2	*("Ridgley", "was in Abrams's apartment that", "night")*
	3	*("Abrams's body", "remains in", "Panama City morgue")*
Berkeley		*The widow of a mortgage executive, Ms. Abrams was something of a force of nature in Staten Island society. The suspect, Debra Ann Ridgley, is.*
SALIE	1	*("Abrams", "had been stabbed to death in", "apartment")*
	2	*("Remains", "were discovered beside warehouse at", "edge of cinder-topped soccer field on outskirts of Panama City")*
	3	*("Apartment", "tending wounds at time of", "murder")*

(a) MINIE *safe* mode.

Human Summary		
Russian state oil company Rosneft has lined up $22 billion in financing from consortium of Western banks to buy assets from bankrupt rival Yukos; Rosneft says it will bid for refineries owned by Yukos as outlet for production from its Yugansk subsidiary in western Siberia; some of banks listed.		
Method		**Salient Facts / Summary**
Position	1	*("State oil company", "lined up $ from consortium of banks buy assets from", "rival")*
	2	*("Rosneft", "increase footprint in", "oil and gas business")*
	3	*("Bids", "are successful as", "expected")*
TextRank	1	*("Banks", "made loans to", "Rosneft and state company")*
	2	*("Banks", "lent company related to Rosneft", "$ increase share")*
	3	*("State oil company", "lined up $ from consortium of banks buy assets from", "rival")*
Berkeley		*The Russian state oil company Rosneft has lined up $22 billion from a consortium of Western banks.*
SALIE	1	*("State oil company", "lined up $ from consortium of banks buy assets from", "rival")*
	2	*("Banks", "made loans to", "Rosneft and state company")*
	3	*("Rosneft", "increase footprint in", "oil and gas business")*

(b) MINIE *aggressive* mode.

Table 1: Top-3 salient facts automatically extracted from a sample of two NYT documents with two different MINIE modes. For Berkeley (which does not return facts) we show its produced summary. On the top of each table we show the summary written by a human for the input document.

2004). ROUGE-1 measures the presence of single words between the salient facts and the summary; ROUGE-L identifies the longest common subsequence (LCS) with maximum length between facts and summary; ROUGE-1.2W measures the weighted LCS by taking into account spatial relations and giving higher values to consecutive matches; ROUGE-SU is the number of occurring bigrams between the facts and summary with arbitrary gaps. For each metric we report the F_1 performance, all computed with a 95% confidence interval, run with stemming and stopword removal[3].

To compute the ROUGE score, the facts were flattened and concatenated into a sequence of tokens respecting the ranking order. For the computation, this sequence is considered equivalent to a summary, so the same conditions apply: If all the extracted tokens fully cover the gold standard summary, the ROUGE score reaches its highest value.

[3]Package arguments: `-c 95 -m -s -U -w 1.2`.

Note that we do not take into account the correctness of the facts (i.e., if they are well-structured). All systems implemented, except the Berkeley summarizer (Durrett et al., 2016), use the same open facts extracted by MINIE. Also for the Berkeley summarizer, we do not evaluate the structure or fluency of the summary.

SALIE. Outputs top-k facts per article. We show results for two MINIE configurations: *safe* and *aggressive*, which differ in the fact average size.

Intrinsic Evaluation. As there is no direct fact salience competitor, we designed three baselines: The standard *Position* baseline which ranks facts with respect to their order of appearance, *Tf-Idf* which ranks them with respect to the subject's head tf-idf and the *Context* baseline which ranks facts with respect to the cosine-similarity between the document and the fact embedding's centroid.

Extrinsic Evaluation. We used two state-of-the-art document summarizers, i.e., the unsupervised graph-based TextRank (Mihalcea and Tarau, 2004) and the supervised Berkeley summa-

Method	ROUGE-1					ROUGE-L					ROUGE-1.2W					ROUGE-SU				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Position	13.9	20.4	24.8	27.8	29.7	12.8	18.1	21.8	24.4	26.0	6.20	8.80	10.8	12.2	13.2	2.70	5.30	7.50	9.00	10.0
Tf-Idf	10.5	15.8	19.0	21.0	22.3	9.60	13.2	15.7	17.5	18.5	4.60	6.30	7.50	8.50	9.20	1.40	2.80	4.00	4.80	5.30
Context	13.6	19.6	22.8	24.6	25.7	11.5	16.3	18.9	20.4	21.3	5.60	8.00	9.40	10.3	11.0	2.50	4.40	5.60	6.30	6.70
TextRank	15.2	21.5	24.5	26.1	26.8	13.0	17.5	19.8	21.3	22.0	6.20	8.40	9.70	10.6	11.2	2.60	4.90	6.40	7.20	7.50
Berkeley	8.50	18.0	25.4	30.4	34.1	8.00	16.3	22.5	26.7	29.7	3.80	7.70	11.0	13.2	14.9	0.80	3.40	6.90	10.1	12.7
SALIE	17.1	24.2	28.0	30.0	30.9	15.3	21.2	24.3	26.0	26.8	7.40	10.3	12.0	13.1	13.6	3.60	6.50	8.30	9.20	9.50
Diff.	**+1.9**	**+2.7**	**+2.6**	-0.4	-3.2	**+2.3**	**+3.1**	**+1.8**	-0.7	-2.9	**+1.2**	**+1.5**	**+1.0**	-0.1	-1.3	**+0.9**	**+1.2**	**+0.8**	-0.9	-3.2

(a) MINIE *safe* mode.

Method	ROUGE-1					ROUGE-L					ROUGE-1.2W					ROUGE-SU				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Position	10.1	15.3	19.3	22.3	24.5	9.60	13.8	17.2	19.8	21.7	4.40	6.30	7.80	9.10	10.1	1.40	2.90	4.50	5.80	6.90
Tf-Idf	8.90	13.6	16.4	18.3	19.6	8.30	11.3	13.4	15.0	16.1	3.80	5.20	6.20	6.90	7.60	0.90	2.00	2.90	3.50	4.10
Context	9.50	14.5	17.9	20.0	21.4	8.40	12.4	15.1	16.8	18.0	3.90	5.70	6.90	7.80	8.50	1.20	2.40	3.40	4.20	4.80
TextRank	11.3	17.2	20.3	22.2	23.3	10.1	14.3	16.7	18.2	19.2	4.60	6.40	7.60	8.40	9.00	1.30	2.90	4.20	5.00	5.60
Berkeley	3.60	10.6	16.2	21.2	25.4	3.50	9.90	14.8	19.0	22.5	1.60	4.60	7.00	9.10	11.0	0.20	1.20	2.80	4.80	6.90
SALIE	11.6	17.9	21.6	24.2	25.9	10.5	15.9	19.1	21.3	22.8	4.80	7.20	8.60	9.70	10.5	1.60	3.30	4.60	5.70	6.50
Diff.	**+0.3**	**+0.7**	**+1.3**	**+1.9**	**+0.5**	**+0.4**	**+1.6**	**+1.9**	**+1.5**	**+0.3**	**+0.2**	**+0.8**	**+0.8**	**+0.6**	-0.5	**+0.2**	**+0.4**	**+0.1**	-0.1	-0.4

(b) MINIE *aggressive* mode.

Table 2: Results on the NYT dataset with two different MINIE modes.

rizer (Durrett et al., 2016). We adapted TextRank to work with facts instead of sentences. For the Berkeley summarizer, we used the model online[4]. As the size of the summaries is a parameter of the summarizer, we set it to match the average size of MINIE facts (*safe* is 10 and *aggressive* is 6), For example, for the top-5 configuration in the aggressive mode, the summary length is set to 30.

Tab. 1 shows example outputs for the position baseline, the text summarizers and SALIE.

5.1 Results

Tabs. 2a and 2b show the results for all the systems and baselines. We use colors **black**, **gray** and **light gray** for the first, second and third best performing methods. In each ROUGE configuration, we show results for five rankings: top-1 to top-5. The difference between SALIE and the best competitor is reported in the last line of the tables.

Tab. 2a shows the results where facts have been extracted with MINIE's *safe* mode. SALIE outperforms all other methods and baselines for the first three rankings (top-1 to top-3), although Berkeley summarizer comes first in top-4 and 5 facts as a higher budget takes the system closer to the gold standard human-readable summaries. TextRank has an opposite behavior compared to Berkeley, performing well in top-1 and 2 but lagging behind as more facts are added probably due to the lack of a diversification stage. It is interesting to note that systems working at the fact level do well in constrained settings, suggesting that facts may be an effective way to compress information.

Tab. 2b shows the results when MINIE is used in *aggressive* mode. In this experiment, we aim to analyze the behavior of the systems in a restricted scenario with a very small budget size (6 tokens per fact). SALIE achieves the highest performance overall metrics independently the number of facts used, with the only exception on the ROUGE-1.2W and ROUGE-SU score when 4 or 5 facts are used. The second and third best performing methods are Position and TextRank. Again, in this case, it is suggested that facts are an appropriate mechanism to compress information.

Overall SALIE shows a more stable balance across all rankings in both settings. It always ranks first or second (except in ROUGE-SU top-5 where it comes third). Compared to TextRank it seems to significantly better manage redundancy, while compared to the Berkeley it does better at detecting relevant information in constrained settings. This is due to the use of facts as a mean to compress information.

6 Conclusions

We introduced the fact salience task. We also presented SALIE, the first fact salience system. SALIE outperformed standard baselines but also state-of-the-art automatic text summarizer. We showed that working at the fact level allows to more effectively compress information.

[4] nlp.cs.berkeley.edu/projects/summarizer.shtml

References

Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum. 2018. Never-ending learning for open-domain question answering over knowledge bases. In *Proceedings of the WWW*.

Mehdi Allahyari, Seyed Amin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. Text summarization techniques: A brief survey. *CoRR*.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of IJCAI*.

Elena Baralis, Luca Cagliero, Saima Jabeen, Alessandro Fiori, and Sajid Shah. 2013. Multi-document summarization based on the Yago ontology. *Expert Systems with Applications*.

Hannah Bast, Florian Bäurle, Björn Buchhold, and Elmar Haussmann. 2014. Semantic full-text search with Broccoli. In *Proceedings of SIGIR*.

Hannah Bast, Björn Buchhold, and Elmar Haussmann. 2017. Overview of the triple scoring task at the WSDM cup 2017. *Proceedings of WSDM*.

Xiao Cheng and Dan Roth. 2013. Relational inference for wikification. In *Proceedings of EMNLP*.

Jen-Tzung Chien and Ying-Lan Chang. 2013. Hierarchical theme and topic model for summarization. In *International Workshop on MLSP*.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2013. Towards coherent multi-document summarization. In *Proceedings of NAACL-HLT*.

Janara Christensen, Stephen Soderland, Gagan Bansal, and Mausam. 2014. Hierarchical summarization: Scaling up multi-document summarization. In *Proceedings of ACL*.

Luciano Del Corro and Rainer Gemulla. 2013. ClausIE: Clause-based open information extraction. In *Proceedings of WWW*.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of ACL*.

Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*.

Kiril Gashteovski, Rainer Gemulla, and Luciano Del Corro. 2017. MinIE: minimizing facts in open information extraction. In *Proceedings of EMNLP*.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of ACL*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of EMNLP*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of IJCNLP*.

Dat Ba Nguyen, Abdalghani Abujabal, Nam Khanh Tran, Martin Theobald, and Gerhard Weikum. 2017. Query-driven on-the-fly knowledge base construction. *Proceedings of VLDB*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*.

Marco Ponza, Paolo Ferragina, and Francesco Piccinno. 2018. Swat: A system for detecting salient wikipedia entities in texts. *CoRR*.

Seyedamin Pouriyeh, Mehdi Allahyari, Krzysztof Kochut, Gong Cheng, and Hamid Reza Arabnia. 2017. Es-lda: Entity summarization using knowledge-based topic modeling. In *Proceedings of IJCNLP*.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *Proceedings of ACL*.

Gabriel Stanovsky, Ido Dagan, et al. 2015. Open ie as an intermediate structure for semantic tasks. In *Proceedings of ACL*.

Shuchu Xiong and Yihui Luo. 2014. A new approach for multi-document summarization based on latent semantic analysis. In *Proceedings of ISCID*.

Libin Yang, Xiaoyan Cai, Yang Zhang, and Peng Shi. 2014. Enhancing sentence-level clustering with ranking-based clustering framework for theme-based summarization. *Information Sciences*.

David Zajic, Bonnie J Dorr, Jimmy Lin, and Richard Schwartz. 2007. Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. *Information Processing & Management*.