Towards Generating Text Summaries for Entity Chains

Shruti Chhabra and Srikanta Bedathur

Indraprastha Institute of Information Technology
New Delhi, India
{shrutic,bedathur}@iiitd.ac.in

Abstract. Given a large knowledge graph, discovering meaningful relationships between a given pair of entities has gained a lot of attention in the recent times. Most existing algorithms focus their attention on identifying one or more structures –such as relationship chains or subgraphs– between the entities. The burden of interpreting these results, after combining with contextual information and description of relationships, lies with the user. In this paper, we present a framework that eases this burden by generating a textual summary which incorporates the context and description of individual (dyadic) relationships, and combines them to generate a ranked list of summaries. We develop a model that captures key properties of a well-written text, such as coherence and information content. We focus our attention on a special class of relationship structures, two-length entity chains, and show that the generated ranked list of summaries have 79% precision at rank-1. Our results demonstrate that the generated summaries are quite useful to users.

Keywords: entity chain, text summarization, relationship queries, entity-relationship graphs

1 Introduction

The use of large entity-relationship graphs such as DBPedia [3], Freebase [5], and Yago [16], in various information retrieval and discovery tasks have given raise to many challenging problems. The problem of discovering long ranging semantic relationships between a pair or a group of entities has attracted much attention recently. Existing solutions take as input a pair of entities and extract structures such as a chain of nodes connecting the input pair [2], or a subgraph [10] from the underlying entity-relationship graph. While these structures are useful in capturing the important relationships, the burden of interpreting the overall relationship lies with the user. The attributes such as the context and relationship descriptions that make such an interpretation possible, are missing from the extracted structure.

Efforts are being made to associate textual evidences (such as documents, sentences, or phrases) that describe underlying relationships in a more human-understandable form. Information extraction systems such as OpenIE [9], PATTY

Table 1: Sample human-generated summaries for two-length entity chains

No.	Entity Chain	Human-generated Summary
1	⟨Brooke Shields, Andre Agassi, Steffi Graf⟩	Andre Agassi was married to Brooke Shields till 1999. He married Steffi Graf in 2001.
2	⟨Richard Nixon, John F. Kennedy, Lyndon Johnson⟩	John F. Kennedy defeated Richard Nixon in 1960 U.S. presidential election. Lyndon Johnson became president after Kennedy.
3	⟨Afghanistan, Opium, Europe⟩	Afghanistan is the world leading producer of Opium. The better quality Opium is smuggled to Europe.
4	⟨Charlie Sheen, Martin Sheen, West Wing⟩	Charlie Sheen is son of Martin Sheen. Martin was a part of television drama, The West Wing.
5	⟨Satwant Singh, Indira Gandhi, Operation Blue Star⟩	Indira Gandhi was killed by two of her sikh bodyguards, Satwant Singh and Beant Singh, in the aftermath of Operation Blue Star. Operation Blue Star was an Indian army's assault on the Golden Temple, ordered by Indira Gandhi.

[25], and NELL [6] routinely maintain textual evidences for relations they extract. Techniques such as support sentence retrieval [4] help to retrieve short passages or sentences from a text corpus for a given entity. Despite these, the complex task of combining these evidences available for individual relationships rests with the user.

To address these issues, we propose the idea of generating textual summaries corresponding to the extracted relationship structures. In this paper, we restrict our attention on simple chain structures consisting of two entities connected through an intermediate entity, which we call as two-length entity chains. Considering sentence-level evidence associated with edges, we develop a model that generates candidate summaries for a given two-length entity chain by combining the sentences associated with individual edges in the chain and ranks these candidate summaries.

Clearly, the summaries generated by the underlying model must satisfy the basic properties of a well-written, human generated summary. Table 1 illustrates some human-generated summaries for sample entity chains. Humans use their knowledge of writing a well formed text in presenting the facts related to given entities, which makes the summaries understandable (as presented in Table 1). The aim of our model is to generate such good quality summaries.

We conducted experiments to identify the key properties of a good quality summary. Based on the insights from literature [1][21][26] and our initial experiments [7], we argue that the following three properties are crucial to capture the intrinsic characteristics of a good summary:

- *Coherence*: Coherence refers to the "sticking together" property of the text. In other words, it is the degree to which pieces of information throughout the text are related and can be linked. For instance, entities Andre Agassi and Steffi Graf in example 1 are connected through two different relationships i.e. Andre Agassi married Steffi Graf and both are tennis players (also have played in same tournaments). However, in the presence of entity Brooke Shields, choosing marriage relationship over tennis makes the summary more coherent.
- *Succinctness*: A summary should be focused and "to the point" – i.e., it should cover only the most important aspect of text. For instance, in example summary 4 above, information such as the mention of Aaron Sorkin, the creator of the television drama West Wing is not desirable.
- *Non-Redundancy*: There should not be redundant information within the summary text. From the model summaries given in Table 1, it is clear that the sentences that form a summary do not contain any redundant information.

The model we have developed that captures these key properties is evaluated on a query set of 27 entity chains. In a preliminary user study, our summaries showed a high precision value of 79% at rank 1. The experimental analysis demonstrates the effectiveness of proposed model and also show that the three properties are helpful in generating good quality summaries.

The layout of the rest of this paper is as follows. We begin with a brief discussion of related work in Section 2. We formally describe our model for generating good summaries for a two-length entity chains in Section 3. We present our experimental framework and describe the results in Section 4, before concluding in Section 5.

2 Related Work

Researchers have been working in various dimensions of the relationship extraction problem such as *relationship queries* and *hypothesis generation* to extract connection between two or more entities. However, the problem of generating summaries for extracted connections is a relatively less explored area of research. We further discuss each dimension in following subsections.

2.1 Relationship Queries

Relationship queries refer to queries posed on entity-relationship graphs to extract connections between two or more entities. Researchers have modeled the relationships among entities in varied ways. Anyanwu and Sheth [2] defined complex relationships on RDF such as paths between entities, networks of paths, or subgraphs as *semantic associations*. Halaschek *et al.* [15] referred the problem of finding paths between entities as *ρ-path semantic associations* and proposed a system called SemDIS to discover such semantic associations in RDF data.

Kasneci *et al.* [20] further extended the problem to find relationships between a set of two or more entities as a Steiner tree computation in weighted entity-relationship graphs. Fang *et al.* [10] proposed a system called REX to find a subgraph in entity-relationship graph which connects the entity pair. They also demonstrated the necessity of including non-simple paths in the results. Srihari *et al.* [30] proposed a framework to generate ranked list of query relevant hypothesis graphs (essentially subgraphs) from concept-association graph.

2.2 Hypothesis Generation

In the mid 1980s, Swanson [32] introduced a closed-discovery framework for hypotheses generation. Given two disconnected topics, they explored MEDLINE to identify potential linkages via intermediate topics and generate interesting hypotheses. These hypotheses were identified by manual analysis. Swanson and Smalheiser [28][29][33] proposed various interesting hypotheses, which were later successfully verified by clinical studies. Srinivasan [31] proposed text mining algorithms to automatically address the problem of closed-discovery algorithms. Researchers have also studied the effectiveness of incorporating domain semantics in the discovery framework [8][17][18]. These ideas have also been imported by the IR community to find associations in web collections [19].

2.3 Summary Generation

Srihari *et al.* [30] proposed a framework to generate hypothesis graphs for two or more entities and ranked evidence trails for the graphs. The evidence trails act as summary for the graph. Jin *et al.* [19] addressed the special case of entity chains. They aim at finding most meaningful evidence trails across documents that connect topics in an entity chain. The technique was tested for topic-specific datasets, namely the 9/11 Commission report and aviation accident reports provided by NTSB, therefore the effectiveness for generic queries is unclear.

3 Entity Chain Summaries

Given a two-length entity chain, we aim to compute a ranked list of summaries by combining the textual evidences, i.e. sentences, associated with the edges in the entity chain. Consider an entity chain represented as $\langle v_1, v_2, v_3 \rangle$, thus the two edges $e_1 = (v_1, v_2)$ and $e_2 = (v_2, v_3)$ are connected by a common node v_2. Each edge e_i is associated with a set of sentences S_{e_i} (e.g., the textual evidences maintained by the information extraction system). Based on this, we can formulate two kinds of candidate summaries – first, which we call as CS_1, is generated by combining individual sentences from the sentence sets of each edge. Formally,

$$CS_1 = \{\ l_1 \oplus l_2 \mid l_1 \in S_{e_1} \cap l_2 \in S_{e_2}\ \},$$

where $\oplus$ denotes concatenation.

The second kind of candidate summary, called CS_2, consists of rich single sentences that contain all the three entities in the entity chain. Formally, we define them as follows:

$$CS_2 = \{\ l_1 \in S_1 \wedge \phi(l_1, v_3)\ \} \cup \{\ l_2 \in S_2 \wedge \phi(l_2, v_1)\ \},$$

where $\phi(l_i, v_j)$ is a indicator function which is true when the entity v_j is mentioned in the sentence l_i.

Finally, the full candidate set of summaries CS is the union of CS_1 and CS_2 as defined above, and consists of summaries with one or two sentences. Now, the task is to rank the elements of CS.

A summary should represent a well-written piece of text, therefore, the ranking function must capture the underlying concepts and structure of a well-written document. Significant research has been made to capture the properties of a well-written text [1][21][26]. Based on these studies and results from our initial experiments [7], we identified three such properties which contribute in enhancing the quality of the summary and make it easier for the reader to comprehend the text. The properties are coherence, succinctness and non-redundancy. The coherence and succinctness properties together are important to form a good quality summary. The lower quotient of any of these may drastically degrade the quality of the summary. Moreover, it is also essential to penalize the summaries which contain redundant information. Based on this rationale, we define following ranking function to score the candidate summaries:

$$Score(m) = (1 - \alpha)(Coherence(m) * Succinctness(m)) - \alpha\ Redundancy(m),$$

where $m \in CS$. The parameter α can be adjusted based on the tolerance level for redundancy. We will now discuss how these three properties are computed in our model.

3.1 Coherence

A coherent text helps the reader to link related pieces of information and comprehend a well connected representation of the text. Various approaches coherence in automatically generated summaries have been proposed in the literature. Referring to the study by Lapata and Barzilay [23], methods based on Latent Semantic Analysis (LSA) are well correlated with human ratings of coherence in text snippets. Various researchers have used LSA to measure coherence [12][13][22]. Moreover, Foltz *et al.* [13] have examined and demonstrated the potential of LSA as psychological model of coherence. At the same time, LSA is well suited for our setting where the summary lengths are quite small.

LSA represents a text corpus as a matrix where each row corresponds to a unique word and each column stands for a sentence. Cells of the matrix contain the frequency of the corresponding word in the sentence. The matrix is then decomposed using Singular Value Decomposition (SVD) such that every sentence is represented as vector whose value is the sum of vectors standing for its component words.

Consider a rectangular matrix X of n terms and m sentences. SVD decomposes this $n \times m$ matrix into product of three matrices as,

$$X = T * S * P' \tag{1}$$

where T and P are the orthogonal matrices and S is a diagonal matrix of eigenvalues. LSA uses a truncated SVD, keeping only k largest eigenvalues and associated vectors. Therefore, $X = T_k * S_k * P_k'$ where T_k and P_k represent the term vectors and sentence vectors in latent semantic space. Similarity between two sentences is directly computed using distance measures on the sentence vectors.

We define *Coherence* score as the cosine similarity between the sentence vectors. Hence, the coherence scores for summaries in set CS_1 are computed as follows,

$$Coherence(l_1 \oplus l_2) = \sum_{i=1}^{k} \frac{ls_1^i . ls_2^i}{(|ls_1^i| . |ls_2^i|)}, \qquad l_1 \oplus l_2 \in CS_1 \tag{2}$$

where ls_1 and ls_2 are the sentence vectors of l_1 and l_2 respectively in the latent semantic space. The coherence scores of single sentence summaries in CS_2 are set to 1.

3.2 Succinctness

Succinctness captures the notion that a more unified and single purpose text results in better comprehension [14]. In our setting, we observed that while entities are the most information dense tokens of text, a text with many entities usually is contextually weaker. Therefore, we consider the number of entities present in the summary as the notion of succinctness. The *Succinctness* score for each summary in CS is computed as,

$$Succinctness(m) = \frac{1}{EntityCount(m)} \tag{3}$$

Succinctness acts as a notion of *relevance* of single sentence summaries for the corresponding entity chain.

3.3 Non-Redundancy

A precise summary should not contain any redundant information. Therefore, it is necessary to focus on the text phrases representing entity relationships in the summary. The presence of redundant relationship phrases across the sentences indicates a low quality summary. In order to handle redundancy, we take a shallow approach by considering a metric which relies on n-gram overlap between the sentences. In our model, the n-gram based similarity of sentences l_1 and l_2 is computed to account for redundancy.

$$Redundancy(l_1, l_2) = \sum_{k=1}^{n} w_k * \frac{|grams(l_1, k) \cap grams(l_2, k)|}{|grams(l_1, k) \cup grams(l_2, k)|} \tag{4}$$

where $(l_1, l_2) \in CS_1$, $grams(s, k)$ is the set of k-grams of sentence s and w_k is the weight associated with k-gram similarity of sentences. We choose $n = 4$ and set $w_1 = 1$, $w_2 = 2$, $w_3 = 3$, and $w_4 = 4$. In case of a single sentence summary, *Redundancy* score is set to 0.

4 Experimental Results

We utilize Open Information Extraction (Open IE) [9] to identify sentence sets for edges between entity pairs. We assume the sentence sets obtained through Open IE convey correct and important relationships. LSA model is learnt on the Wikipedia corpus (published on Sep 13, 2013) containing about 4.2 million documents using Gensim [27], an open source topic modeling tool. The number of topics to be learnt is set to 400. The corpus is annotated with named entities using Stanford's CoreNLP kit [11]. The tolerance parameter for redundancy (α) in the ranking function is empirically set to 0.2.

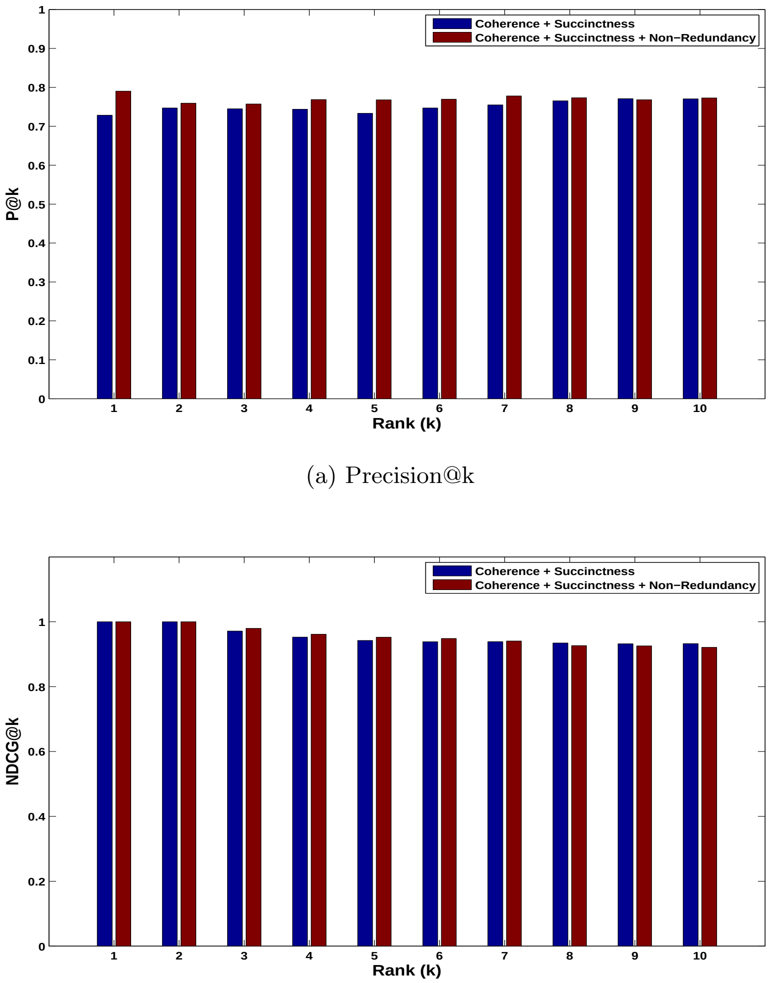
We manually constructed a set of 27 entity chains for the purpose of evaluations. We select only those entity chains for which OpenIE provides non-empty sentence sets. The query set was designed to contain different types of entities such as person, organization, date, product, and country.

Four different summary ranking models are possible based on the following combinations of properties we proposed: (i) only coherence, (ii) coherence and succinctness, (iii) coherence and non-redundancy, and (iv) coherence, succinctness, and non-redundancy. However, we use only (ii) and (iv) in our analysis since the notion of relevance for single sentence summaries does not exist in the case of models not considering succinctness.

Top-10 output summaries are considered for user evaluation. The summaries are evaluated by three judges, who are asked to assign a grade to each summary using four levels: 1 for poor, 2 for average, 3 for good, and 4 for perfect. Following definition of these levels were given to judges prior to evaluations:

- **Perfect** when the relationships are explicitly mentioned and are topic of discussion,
- **Good** when the relationships are explicitly mentioned but they are not topic of discussion,
- **Average** when the relationships are not mentioned but can be inferred, and
- **Poor** when the relationships are neither mentioned nor can be inferred.

The performance measures – P@k and NDCG@k are calculated. While computing P@k, summaries graded as average or higher are marked as relevant. Note that in literature, ROUGE [24] family of measures is most frequently used for evaluating automatic summarization results. However, appropriate ROUGE parameterization varies with task at hand. Since the problem of generating summaries for a given entity chain has not been addressed before to the best of our knowledge, an extensive experimental study is needed before selecting the appropriate ROUGE measure. While we pursue this research direction further, in this paper we settle for an evaluation based on precision and NDCG metrics.



(a) Precision@k

(b) NDCG@k

Fig. 1: Precision and NDCG of output summaries.

Figure 1 compares the performance of the ranking model that considers only the coherence and succinctness of summaries, with the model that also includes non-redundancy defined above. In figure 1a, the average precision of summaries at different rank positions are shown. It is worth observing that the overall precision of the resulting summaries is consistently above 70% with both methods. The inclusion of non-redundancy in the model improves the precision of the result in rank-1 by more than 5%, suggesting the importance of non-redundancy

Table 2: Top-ranked system-generated summaries for entity chains from Table 1.

No.	Entity Chain	Summary
1	⟨Brooke Shields, Andre Agassi, Steffi Graf⟩	Agassi, who was previously married to actress Brooke Shields, married Steffi Graf in 2001.
2	⟨Richard Nixon, John F. Kennedy, Lyndon Johnson⟩	Campaign Issues Ask students the following: Why did John F. Kennedy choose Lyndon Johnson as his vice president? The Eisenhower years were petering out and Nixon was running against Kennedy who won because things were not going so well at all.
3	⟨Afghanistan, Opium, Europe⟩	According to UNODCCP, in recent years Afghanistan had been the main source of illicit opium: 70 percent of global illicit opium production in 2000 and up to 90 percent of heroin in European drug markets originated from the country.
4	⟨Charlie Sheen, Martin Sheen, West Wing⟩	Sheen is the son of actor Martin Sheen, former star of the hit TV drama The West Wing.
5	⟨Satwant Singh, Indira Gandhi, Operation Blue Star⟩	Hindu men rampage through the streets of Delhi, November 1984 On October 31, 1984, the Indian Prime Minister, Indira Gandhi, who had ordered Operation Bluestar, was assassinated in a revenge attack by her two Sikh bodyguards. Ever wonder why Sonia had asked the President of India to set aside on a mercy petition the Supreme Court judgment directing that Rajiv Gandhis LTTE killers be hanged, more particularly, when she had not similarly acted for Satwant Singh who killed Indira Gandhi?

property in enhancing the quality of summary. Further, the precision of the best performing model at rank-1 is 79%, which show that the model is capable of automatically constructing a good summary for a given entity chain. Similarly, NDCG@k plotted in figure 1b for both the models is more than 92% across all the top-10 ranks. This further gives strength to our claim that the ranking of summaries are consistently accurate in being able to clearly explain the relationships in the given entity-chain.

Table 2 illustrates the top summary retrieved for the entity chains illustrated in Table 1 using the model that combines all the three properties we have considered. Following qualitative aspects can be inferred from the results illustrated in Table 2:

- On comparison with expected summaries presented in Table 1, the sentences in output summaries are topically similar, this illustrates the effectiveness of LSA as a coherence measure.

- Single sentence summaries are usually more coherent than two sentence summaries. Thus, it validates our assertion that single sentence summaries should be included in the candidate set.
- Summaries with lesser entities are more focused and to the point. It confirms the usefulness of incorporating succinctness property in the model.
- The summary generated for entity chain 2 weakens our assumption that the sentences from Open IE are correctly describing the relationship.

The results presented show that the proposed system effectively generates summary for an entity chain with a precision of 79% at rank 1.

4.1 Discussion

The performance of proposed framework is majorly affected by two parameters- the quality of sentence sets obtained from Open IE and evaluation dataset. In this section, we discuss the challenges associated with each of these parameters.

The quality of our results are highly dependent on the results produced by the framework generating sentence sets. As discussed above, there are various systems available such as NELL [6] and Patty [25] to extract relationship between entities. Open IE is one such system which provides list of sentences associated with the extracted relationships. But, Open IE is still restricted by the variety in type of relationships it can extract from unstructured text. Moreover, a methodology to generate sentence sets aligned more towards the task of summary generation may help in enhancing the quality of summaries. Since our work is one of the initial steps towards generating summaries for graph snippets, there is no open evaluation dataset available. As discussed above, the selection of 27 entity chains in our evaluation dataset is restricted by the output from Open IE.

5 Conclusion and Future Work

This research address the challenging problem of generating textual summaries for the two-length entity chains. The proposed system is build upon the key characteristics of a well written document. The results show that summaries generated by the system enable user to understand the underlying relationships. A high precision of 79% at rank 1 shows promise for more in-depth work.

The work presented here can be extended along various dimensions such as generating summary for generic entity chains, knowledge subgraphs. The proposed model can be augmented with other important properties of a well-written text pertaining to the problem in hand. A deeper analysis of sentences in sentence sets may help in filtering non-declarative sentences, thus enhancing quality of the summary. Further, entities in an entity pair may be connected to each other through various relationship categories such as personal, political, and professional. The ranked summaries can be diversified based on these categories so that wide range of information about the relationship can be covered.

References

1. R. Agrawal, S. Chakraborty, S. Gollapudi, A. Kannan, and K. Kenthapadi. Empowering authors to diagnose comprehension burden in textbooks. In *KDD*, pages 967–975, 2012.
2. K. Anyanwu and A. Sheth. The ρ operator: discovering and ranking associations on the semantic web. *ACM SIGMOD Record*, 31(4):42–47, 2002.
3. C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. Dbpedia - a crystallization point for the web of data. *Web Semant.*, 7(3):154–165, 2009.
4. R. Blanco and H. Zaragoza. Finding support sentences for entities. In *SIGIR*, pages 339–346, 2010.
5. K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *SIGMOD*, pages 1247–1250, 2008.
6. A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell. Toward an architecture for never-ending language learning. In *AAAI Conf. on Artifical Intelligence*, 2010.
7. S. Chhabra and S. Bedathur. Generating text summaries of graph snippets. In *COMAD*, pages 121–124, 2013.
8. T. Cohen, G. Whitfield, R. Schvaneveldt, K. Mukund, and T. Rindflesch. Epiphanet: An interactive tool to support biomedical discoveries. *Journal of Biomedical Discovery and Collaboration*, 5:21–49, 2010.
9. O. Etzioni, A. Fader, J. Christensen, S. Soderland, and M. Mausam. Open information extraction: The second generation. In *IJCAI*, pages 3–10, 2011.
10. L. Fang, A. D. Sarma, C. Yu, and P. Bohannon. Rex: explaining relationships between entity pairs. *Proc. VLDB Endow.*, 5(3), 2011.
11. J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, pages 363–370, 2005.
12. P. W. Foltz. Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, & Computers*, 28(2):197–202, 1996.
13. P. W. Foltz, W. Kintsch, and T. K. Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307, 1998.
14. W. S. Gray and B. E. Leary. *What makes a book readable.* Univ. Chicago Press, 1935.
15. C. Halaschek, B. Aleman-Meza, I. B. Arpinar, and A. P. Sheth. Discovering and ranking semantic associations over a large rdf metabase. In *VLDB*, pages 1317–1320, 2004.
16. J. Hoffart, F. M. Suchanek, K. Berberich, E. Lewis-Kelham, G. de Melo, and G. Weikum. Yago2: Exploring and querying world knowledge in time, space, context, and many languages. In *WWW*, pages 229–232, 2011.
17. D. Hristovski, C. Friedman, T. C. Rindflesch, and B. Peterlin. Exploiting semantic relations for literature-based discovery. In *AMIA Annual Symp.*, volume 2006, pages 349–353, 2006.
18. D. Hristovski, A. Kastrin, B. Peterlin, and T. C. Rindflesch. Combining semantic relations and dna microarray data for novel hypotheses generation. In *Proceedings of the 2009 Workshop of the BioLink Special Interest Group, International Conference on Linking Literature, Information, and Knowledge for Biology*, pages 53–61, 2010.

19. W. Jin, R. K. Srihari, H. H. Ho, and X. Wu. Improving knowledge discovery in document collections through combining text retrieval and link analysis techniques. In *ICDM*, pages 193–202, 2007.
20. G. Kasneci, M. Ramanath, M. Sozio, F. M. Suchanek, and G. Weikum. Star: Steiner-tree approximation in relationship graphs. In *ICDE*, pages 868–879, 2009.
21. W. Kintsch and T. A. Van Dijk. Toward a model of text comprehension and production. *Psychological review*, 85(5):363–394, 1978.
22. T. K. L. D. Laham and P. Foltz. Learning human-like knowledge by singular value decomposition: A progress report. In *NIPS*, volume 10, pages 45–51, 1998.
23. M. Lapata and R. Barzilay. Automatic evaluation of text coherence: Models and representations. In *IJCAI*, pages 1085–1090, 2005.
24. C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, 2004.
25. N. Nakashole, G. Weikum, and F. Suchanek. Patty: a taxonomy of relational patterns with semantic types. In *EMNLP*, pages 1135–1145, 2012.
26. E. Pitler and A. Nenkova. Revisiting readability: a unified framework for predicting text quality. In *EMNLP*, pages 186–195, 2008.
27. R. Řehůřek and P. Sojka. Software framework for topic modelling with large corpora. In *LREC Workshop on New Challenges for NLP Frameworks*, pages 45–50, 2010.
28. N. R. Smalheiser and D. R. Swanson. Indomethacin and alzheimer's disease. *Neurology*, 46(2):583–583, 1996.
29. N. R. Smalheiser and D. R. Swanson. Linking estrogen to alzheimer's disease an informatics approach. *Neurology*, 47(3):809–810, 1996.
30. R. K. Srihari, L. Xu, and T. Saxena. Use of ranked cross document evidence trails for hypothesis generation. In *KDD*, pages 677–686, 2007.
31. P. Srinivasan. Text mining: generating hypotheses from medline. *Journal of American Society for Information Science and Technology*, 55(5):396–413, 2004.
32. D. R. Swanson. Two medical literatures that are logically but not bibliographically connected. *Journal of the American Society for Information Science*, 38(4):228–233, 1987.
33. D. R. Swanson and N. R. Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial intelligence*, 91(2):183–203, 1997.