# Utilizing Graph Measure to Deduce Omitted Entities in Paragraphs

**Eun-kyung Kim, Kijong Han, Jiho Kim, Key-Sun Choi**
Semantic Web Research Center
Korea Advanced Institute of Science and Technology (KAIST)
Republic of Korea
{kekeeo,han0ah,hogajiho,kschoi}@kaist.ac.kr

## Abstract

This demo deals with the problem of capturing omitted arguments in relation extraction given a proper knowledge base for entities of interest. We introduce the concept of a salient entity and use this information to deduce omitted entities in the paragraph which allows improving the relation extraction quality. The main idea to compute salient entities is to construct a graph on the given information (by identifying the entities but without parsing it), rank it with standard graph measures and embed it in the context of the sentences.

## 1 Introduction

As the need for structured knowledge for a variety of applications such as knowledge base (KB) completion (Socher et al., 2013), search (Marco and Navigli, 2013), and question-answering (Yahya et al., 2012) has increased, there has been considerable interest in extracting relationships for a large number of documents written in natural language. Relation extraction aims to identify and recognize the semantic relationships between pairs of entities (persons, locations, organizations, etc.) from sentences written in free text and to create them in a structured form.

Most studies in relationship extraction are distantly supervised and only take into account intra-sentence relationships that contain pairs of entities (Mintz et al., 2009; Fan et al., 2014; Zeng et al., 2015). For example, suppose that the following paragraph is given with entities marked by parentheses:"*[Cristiano Ronaldo] was born in Madeira. He plays for the Spanish club [Real Madrid C.F.] and the position is a [Forward].*" Although the entity mentions do not occur in the same sentence, these sentences convey the team to which "Cristiano Ronaldo" belongs and his position, but this cannot be inferred from each individual sentence. In particular, it is very common for an entity to be omitted from a sentence in Wikipedia–a popular corpus for relation extraction–because Wikipedia pages each focus on only one entity in most cases. This is also a very common phenomenon in text that is written in a language that can omit a subject or object even if it is not a Wikipedia article.

There have been studies into tackling these constraints on relation extraction in two or more sentences (Peng et al., 2017; Quirk and Poon, 2017); these are basically done in a way that increases the number of possible paths between the entities present in other sentences by integrating dependency graphs generated in a single sentence. The dependency graph–the key element of these studies–is known to be effective in relation extraction. However, it is difficult to acquire a highly efficient parser for all languages; thus, the practical application cannot extract relationships in various language environments. As another solution, we can apply a pipelined model to first perform a co-reference resolution (Clark and Manning, 2015) or zero-anaphora resolution (Mitkov, 1999) and then perform relation extraction, but error propagation between processes has been pointed to as a common problem in many natural language processing (NLP) tasks (Quirk and Corston-Oliver, 2006; Yang and Cardie, 2013; Han et al., 2013; Zeng et al., 2015).

This demo aims to overcome these issues by means of a projection in the context of the paragraph into the relationship between tuples in the KB. A paragraph is a series of sentences that fleshes out a coherent theme and maintains a consistent flow, so if an omitted entity exists, it is clear that the reader can





*Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 48–52
Santa Fe, New Mexico, USA, August 20-26, 2018.

recognize it as an aspect of the subject that will continue being discussed in the paragraph. Therefore, our assumption here is that we can create a coherent graph composed of nodes (i.e. KB entities) and edges (i.e. relationships between entities) in the paragraph. However, in the conventional distant supervision paradigm, entities from the imperfect sentences in a contiguous context will be unreachable. The key to our approach is to first find the most "salient entity" through KB-based graph interpretation without syntactic parsing or other NLP tools and to maximally associate this with unreachable entities in the paragraph.

## 2 Salient Entity Detection

Normally, a paragraph that consists of a group of sentences deals with a coherent topic, so any reference can be omitted as long as the context provides the subject to which it is referring. In particular, subjects or objects are often omitted when they are obvious from the context. This paper attempts to deal with the null-subject problem to process relation extraction beyond the sentence level; since the subjectivity of an entity can be determined by how it is presented in a paragraph, the "**salience**" of an entity can be computed effectively from what is available in the paragraph itself.

We observed certain cues when identifying salience. Unsurprisingly, salient entities tend to be mentioned in the title or first sentence and are mentioned frequently throughout. However, being included in the title (or first sentence) is neither necessary nor a sufficient condition for salience. Based on these observations, we believe that a KB-based projection of a paragraph that already contains a variety of evidence for an entity is better than developing simple heuristics. This paper defines salient entities as those that have a major impact on the cohesion that occurs in a graph. This assumption is not arbitrary; some of these regularities have been recognized in Centering Theory (Walker et al., 1998). With this goal in mind, we propose a mathematical model and an algorithm to maximize the total connectivity in this situation.

### 2.1 Task Definition

Let $\mathbf{P}$ and $\mathbf{E}$ be the sets of all paragraphs in a given corpus and the set of all entities in the given KB respectively. Let $\mathbf{E}_p \subset \mathbf{E}$ be the set of entities mentioned in $p \in \mathbf{P}$. We formally define the salient task as learning the function:

$$\sigma : \mathbf{P} \times \mathbf{E} \rightarrow \mathbf{R}, \tag{1}$$

where $\sigma(p,e)$ reflects the salience of $e$ in $p$. We denote the ranking of $\mathbf{E}_p$ according to $\sigma$ as:

$$\mathbf{x}_p = \left(e_1, ..., e_{|\mathbf{E}_p|} | e_i \in \mathbf{E}_p, \sigma(p,e_i) \geq \sigma(p,e_{i+1})\right), \tag{2}$$

where pairs of entities with tied scores are arbitrarily ordered.

Our ranking function maximizes coherence in the paragraph-driven-graph by adding outgoing edges from the salient entity to other entities. Maximizing cohesion means creating a maximally connected graph that has the minimum number of entities whose deletion from $\mathsf{G} = (\mathbf{E}_p, \mathbf{A})$ results in a disconnected or trivial graph, where $\mathbf{A}$ is a set of ordered pairs of entities $(e_i, e_j)$. There are two conditions that constitute $\mathbf{A}$: First, $e_i$ and $e_j$ are in a single sentence; second, $e_i$ is a salient entity and $e_j \in \mathbf{E}_p$. Our objective function is expressed as follows:

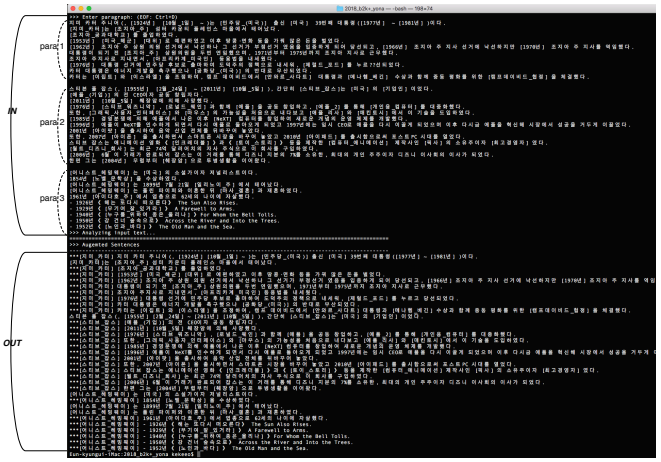
$$\kappa(\mathsf{G}) = \kappa((\mathbf{E}_p, \mathbf{A})) = \sum_{i=1}^{|\mathbf{E}_p|} \sum_{j=1}^{|\mathbf{E}_p|} y_{ij}, \quad y_{ij} = \begin{cases} w_{ij}, & \text{if}(i,j) \in \mathbf{A} \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

where $w_{ij}$ represents the number of relations (i.e. weight) associated with $(i, j)$.

## 3 Evaluation

### 3.1 Experimental Setup

Our experiments aimed to answer whether the artificially restored sentences create noise in the existing distant supervision model, and whether the deduced entities accurately determine more relationships




(a) Sentence Augmentation

지미_카터country 조지아_주. 0.099
스티브_잡스 deathPlace 미국 . 0.026
어니스트_헤밍웨이 deathPlace 미국 . 0.391
어니스트_헤밍웨이 birthPlace 일리노이_주 . 0.054
어니스트_헤밍웨이 notableWork 무기여_잘_있거라 . 0.801

(b) Extracted triples (conventional DS paradigm)

지미_카터successor 1977년 . 0.01
지미_카터birthPlace 미국 . 0.666
지미_카터country 조지아_주. 0.099
지미_카터predecessor 1970년 . 0.061
지미_카터birthPlace 1962년 . 0.737
지미_카터birthPlace 조지아_주. 0.586
지미_카터successor 제럴드_포드 . 0.063
지미_카터party 공화당_(미국) . 0.049
지미_카터country 이스라엘 . 0.656
지미_카터successor 메나헴_베긴 . 0.012
지미_카터birthPlace 이집트 . 0.749
지미_카터successor 안와르_사다트 . 0.016
스티브_잡스 deathPlace 미국 . 0.026
스티브_잡스 occupation 애플_(기업) . 0.764
스티브_잡스 spouse 1996년 . 0.034
스티브_잡스 occupation 최고경영자 . 0.241
스티브_잡스 spouse 2006년 . 0.11
스티브_잡스 occupation 2004년 . 0.295
어니스트_헤밍웨이 deathPlace 미국 . 0.391
어니스트_헤밍웨이 birthPlace 일리노이_주 . 0.054
어니스트_헤밍웨이 deathPlace 아이다호_주 . 0.24
어니스트_헤밍웨이 notableWork 무기여_잘_있거라 . 0.801
어니스트_헤밍웨이 notableWork 누구를_위하여_종은_울리나 .0.863
어니스트_헤밍웨이 notableWork 노인과_바다 . 0.87

(c) Extracted triples by deducing omitted entities

Figure 1: (a) shows the output after restoring the omitted entity from the input sentences. The restored sentence includes the '***' symbol at the front of each sentence. (b) and (c) show the result of the relation extraction, and the result of using the restored sentence in the step (a), respectively.

from the concealed paragraphs. We conducted experiments on the relation extraction between DBpedia entities in a null-subject language Wikipedia (i.e. Korean). We conducted an experiment performing training and testing using the Korean versions of Wikipedia (dumps on July 2017)[1] as the textual corpus source. We chose the dump of Korean DBpedia KB[2] as the background resource. In this experiment's first stage, we transform Wikipedia's links into entity annotations, and the original sentences of the given corpus can thus be automatically annotated with DBpedia entities. We converted each sentence into a word-level matrix in which each row was a sentence vector extracted from our model. Sentence vectors were learned from the Distributed Memory version of the Paragraph Vector algorithm using training data to automatically learn and predict corresponding relationships by the multi-class logistic regression classifier into one of the 50 relation types in our evaluation dataset.

In the real dataset, the whole labeled sentences have an imbalance in the number of labeled relation types. We found that approximately 85% of relations (of total of 215 relations) have fewer than 1,000 instances, and the amount of data in the top 50 relationships is greater than the rest of the data. Hence, we conducted a relation classification for the top 50 relationships except for those that have very little labeled data. There is no gold annotated dataset under distant supervision, so evaluation typically uses the held-out strategy. A held-out evaluation has the advantage of being automatic, but it can produce biased results because a pair of entities known to have no relationship may actually have a relationship. We solved this problem by creating a gold standard that eliminates false negatives by evaluating people. For this, ten college students judged true or false for the noisy gold test-data generated by the distant supervision assumption[3]. We obtained the precision, recall, and F1-scores for each of the 50 relation types in the experiment then the sum of the weighted averages for each performance measure from each class.

We developed the system to verify the approach to salient entity identification in the experiment as shown in Figure 1. The experimental results show that the effectiveness result of creating large volumes of additional training data to learn the KB relation by obtaining missing entities in relation extraction.

### 3.2 Result Analysis: Salient Entity Detection Techniques

Table 1 shows the experimental result for our model ($\mathscr{A}$**(Centrality)**) with various competitors to measure the saliency of the entity for the gold test data. For example, other plausible ways to detect saliency are (1) the entity corresponding to the Wikipedia page ($\mathscr{A}$**(Title)**), (2) the most frequent entity in the

[1] https://dumps.wikimedia.org/kowiki/

[2] http://downloads.dbpedia.org/2016-10/core-i18n/ko/mappingbased_objects_ko.ttl.bz2

[3] All data used in this experiment are provided in: `https://github.com/kekeeo/SASE`

Table 1: The results of experiments with various baselines for saliency.

| | Precision | Recall | F1-Score |
|---|---|---|---|
| $\mathscr{A}$(Centrality) | **0.58** | **0.54** | **0.52** |
| $\mathscr{A}$(Title) | 0.47 | 0.42 | 0.38 |
| $\mathscr{A}$(Max) | 0.52 | 0.48 | 0.45 |
| $\mathscr{A}$(First) | 0.51 | 0.46 | 0.43 |
| Standard | 0.44 | 0.40 | 0.38 |

paragraph ($\mathscr{A}$**(Max)**), (3) the first entity in the paragraph ($\mathscr{A}$**(First)**). The conventional distant supervised relation extraction corresponds to a single sentence that contains two entities (**Standard**), but we augmented this to tasks for two entities in a paragraph as described in above.

As shown in Table 1, since the method of sentence augmentation by adding the omitted entity to the sentences is higher than the conventional paradigm (i.e. **Standard**), we can see that the proposed sentence augmentation method has increased the positive learning instances for relation extraction. Although the method using centrality obtains superior performance than other heuristic methods, it can be seen that incorrect augmented sentences do not positively affect relation extraction, as shown in the comparative performance between $\mathscr{A}$**(Title)** and **Standard**.

## 4 Conclusion

This paper demonstrates a method of learning useful context features necessary to classify relations efficiently in a language environment that features frequent subject omissions and a high density of sentences with imperfect sentence components. Our approach provides a simple yet effective method to incorporate paragraph-level information through capturing missing relation argument model. This is the first distant supervision approach that resolves the problem of data sparseness by alleviating distant supervision assumptions for the relation classification of incomplete sentences to the best of our knowledge. This method has promising potential applications in languages that lack advanced NLP tools.

## Acknowledgments

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government (MSIT) (2013-0-00109, WiseKB: Big data based self-evolving knowledge base and reasoning platform).

## References

Kevin Clark and Christopher D. Manning. 2015. Entity-centric coreference resolution with model stacking. In *ACL (1)*, pages 1405–1415. The Association for Computer Linguistics.

Miao Fan, Deli Zhao, Qiang Zhou, Zhiyuan Liu, Thomas Fang Zheng, and Edward Y. Chang. 2014. Distant supervision for relation extraction with matrix completion. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 839–849, Baltimore, Maryland, June. Association for Computational Linguistics.

Dan Han, Pascual Martínez-Gómez, Yusuke Miyao, Katsuhito Sudoh, and Masaaki Nagata. 2013. Effects of parsing errors on pre-reordering performance for chinese-to-japanese smt. In *PACLIC*. National Chengchi University, Taiwan.

Antonio Di Marco and Roberto Navigli. 2013. Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, 39(3):709–754.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ruslan Mitkov. 1999. Anaphora resolution: The state of the art. Technical report.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics*, 5:101–115.

Chris Quirk and Simon Corston-Oliver. 2006. The impact of parse quality on syntactically-informed statistical machine translation. In Dan Jurafsky and Éric Gaussier, editors, *EMNLP*, pages 62–69. ACL.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *EACL (1)*, pages 1171–1182. Association for Computational Linguistics.

Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *NIPS*, pages 926–934.

Marilyn Walker, Aravind K. Joshi, and Ellen F. Prince, editors. 1998. *Centering Theory in Discourse*. Clarendon Press, Oxford.

Mohamed Yahya, Klaus Berberich, Shady Elbassuoni, Maya Ramanath, Volker Tresp, and Gerhard Weikum. 2012. Natural language questions for the web of data. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 379–390, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bishan Yang and Claire Cardie. 2013. Joint inference for fine-grained opinion extraction. In *ACL (1)*, pages 1640–1649. The Association for Computer Linguistics.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 1753–1762. The Association for Computational Linguistics.