Learning to Extract Comparison Points of Entity Pairs from Wikipedia Articles

Sandeep Kumar Pani
Indian Institute of technology, Kharagpur
sandeepkumarpani888@gmail.com

Naresh R
Indian Institute of technology, Kharagpur
nareshmdu@gmail.com

Pawan Goyal
Indian Institute of technology, Kharagpur
pawang@cse.iitkgp.ernet.in

Plaban Kumar Bhowmick
Indian Institute of technology, Kharagpur
plaban@gmail.com

ABSTRACT

In this paper, we present preliminary results on a novel task of extracting comparison points for a pair of entities from the text articles describing them. The task is challenging as comparison points in a typical pair of articles tend to be sparse. We presented a multi-level document analysis (viz. document, paragraph and sentence level) for extracting the comparisons. For extracting sentence level comparisons, which is the hardest task among three, we have used Convolutional Neural Network (CNN) with features extracted around ⟨ entity, aspect, value ⟩ triple. Experiments conducted on a small dataset provide encouraging performance.

CCS CONCEPTS

• **Computing methodologies → Information extraction**;

KEYWORDS

Comparison extraction; Entity comparison; Information Extraction

ACM Reference Format:
Sandeep Kumar Pani, Naresh R, Pawan Goyal, and Plaban Kumar Bhowmick. 2018. Learning to Extract Comparison Points of Entity Pairs from Wikipedia Articles. In *JCDL '18: The 18th ACM/IEEE Joint Conference on Digital Libraries, June 3–7, 2018, Fort Worth, TX, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3197026.3203909

1 INTRODUCTION

Automatic extraction of metadata of several forms has become one of the primary research areas in digital library. While these systems focus on extracting structured information about individual resources, extraction of information that relates multiple resources together is generally overlooked. Comparison of entities present in multiple resources is one such example. Comparisons have immense value by providing the readers concise view of differences, pros and cons of the entities or processes in focus. It is very hard to imagine book chapters in different disciplines without comparisons. Among many tasks, the researchers in different domains have to perform comparative analysis of entities and methods by analyzing multiple articles. Previous work [6] in this area uses seeded topics to get comparisons. However, extracting structured comparisons automatically from free text is an extremely challenging task and has not been explored much. There have been efforts to extract comparisons of consumer goods from product reviews [4]. The product reviews tend to be short and are very precise about different attributes of the products. Hence, these approaches are not suitable for comparison extraction from larger text body (e.g., encyclopedic description like Wikipedia articles, scientific articles) where the comparison points are immensely sparse. Furthermore, in comparison to the product reviews, the comparable descriptions in larger text body are camouflaged by the nuances of different surface representations of similar information.

In this work, our objective is to extract comparison points from a pair of Wikipedia articles describing two entities. In this respect, our paper makes the following key contributions:

- We define a novel task of generating comparison points of two entities from the respective unstructured data.
- We present a working pipeline that generates such a comparison from a large corpus without any external seed.
- We present a performance analysis of comparison extraction system with respect to a human annotated dataset.

2 NOTION OF COMPARISON

A comparison refers to a consideration or estimate of the similarities or dissimilarities between two entities, where an 'entity' refers to the main subject of the text. The 'entity' is assumed to be the subject of discussion in each sentence, but exceptions do exist. 'Aspect' refers to topic being discussed in a sentence. 'Value' describes the attributes/characteristics of the entity with respect to the aspect of the sentence. To illustrate the definitions further, we use the sentence "*Lion's prey consists of primarily mammals.*". The breakdown of this sentence would be: Entity: *Lion*, Aspect: *prey*, Value: *mammals*.

We define two sentences to be 'comparable' if

- the **focus** entities are different in semantic space but not at very large distance to be suitable objects for comparison (e.g. algae-bacteria vs. algae-mammal)
- the **aspects** are same or semantically similar (e.g., hunting, reproduction habit etc.)
- the **values** that the respective aspects assume are different (e.g., 'savannah' and 'tundra' make for good values for comparison when comparing on the attribute 'habitat'.)

JCDL '18, June 3–7, 2018, Fort Worth, TX, USA

ACM ISBN 978-1-4503-5178-2/18/06.
https://doi.org/10.1145/3197026.3203909

Table 1: Two examples of ⟨ entity, aspect, value ⟩ extraction

Sentence	Entity	Aspect	Value
1. Lion's prey consists of primarily mammals.	Lion	prey	mammals
2. Seeds and fruit form the major component of the diets of pigeons and doves.	pigeons	diets	Seeds and fruit

i.e., the comparable sentences for the two entities mention the exact same aspect (i.e., context), but can be differentiated on the values. A positive example for a comparable pair of sentences, as per our definition, can be "*Llamas typically live for 15 to 25 years, with some individuals surviving 30 years or more.*" and "*The average life expectancy of a camel is 40 to 50 years.*" A negative example for comparable pair of sentences is: "*Most baboons live in hierarchical troops.*" and "*Gorillas have a patchy distribution.*"

3 COMPARISON EXTRACTION PIPELINE

We use a pair of Wikipedia articles as our input. Our algorithm can be divided into 3 parts for deciding comparability at i) the document level, ii) paragraph level and iii) sentence level. We currently restrict our domain to biology to obtain a lexical tree with relations between all possible entities and use graph heuristics to establish comparability between documents. For such comparable entities, we use concept extraction to obtain the topic of each paragraph and then WordMover's distance [5] & Latent Semantic Indexing (LSI) to establish comparability at the paragraph level.

For a dataset consisting of 669 number of Wikipedia documents, the document level precision and recall values are 87.5% and 77.7%. For paragraph level comparability extraction, the performance of the system is 41.66% recall and 45% precision based on 836 number of paragraph pairs in the dataset. We used the response from sentence level annotations to mark the paragraph and document pairs to be comparable.

4 SENTENCE LEVEL COMPARISON

We used Open IE[1] to get the sentence structure, i.e., to extract the ⟨ entity, aspect, value ⟩ in a sentence. We then used a CNN (Convolutional Neural Network) with the following features (denoted by labels):

- F1 OpenIE promises to extract the relationship between entity pairs in a sentence. This feature describes how far the verb and aspect are in the sentence.
- F2 Wu-Palmer metric from Wordnet[2] to get how different the aspects between a pair of sentence were.
- F3 Path-similarity from Wordnet to estimate distance of the values.
- F4 We also convoluted the PMI (Pointwise Mutual Information) [1] values between the aspect and value of one sentence with another to encode how much value the 'aspects' add to the comparability of the sentences in the context of the given 'values'.

- For multi-perspective CNN [3], our input was a 0-1 classification label of whether or not 2 sentences are comparable for the training dataset. The vocabulary was built with the STS dataset[3] added to words from our training dataset.
- We used Infersent [2] for getting a sentence level representation, as it scaled well for different tasks in the NLP domain.

[1]Stanford Open Information Extraction

[2]WordNet - A Lexical Database for English

Table 2: Results for sentence-level comparison extraction

Method	Prec.	Recall	F1
- Open IE + graph similarity	0.286	0.32	0.302
- CNN (with infersent)	0.367	0.393	0.379
- Multi-perspective CNN	0.458	0.423	0.44
- CNN (with features F1,F2,F3; tuples extracted by hand)	0.461	**0.562**	0.507
- CNN (with features F2, F3; triples extracted by hand)	0.3	0.346	0.321
- CNN (features F1,F2,F3,F4; triples extracted by hand)	**0.476**	0.555	**0.513**

5 EXPERIMENTAL RESULTS

We used a human annotated dataset of 300 document pairs. The average number of comparable sentences per document is 3. This exhibits the difficulty of the task. Our final training dataset is comprised of 168 positive pairs and 264 negative pairs. Similarly, test data has 20 positive and 54 negative pairs. In both the cases, positive refers to the sentences that are labeled as comparable. Table 2 presents the results of various approaches explored in this study. It is evident from the results that adding more refined features which highlight the comparability aspect for the pair of sentences is very important. Increase in performance has been observed by using PMI-based features. For training as well as testing, we hand broke the sentence into ⟨ entity, aspect, value ⟩ and this has been reported in the table as 'triples extracted by hand'.

6 FUTURE WORK

The present system makes use of OpenIE for extraction of ⟨ entity, aspect, value ⟩. Though it works well for simpler sentences, quality of extraction has been poor for complex sentences. We intend to deploy template-based extraction of ⟨ entity, aspect, value ⟩ triples.

7 ACKNOWLEDGEMENT

This work is partially supported by *Development of National Digital Library of India as a National Knowledge Asset of the Nation* sponsored by Ministry of Human Resource Development, Government of India.

REFERENCES

[1] Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL* (2009), 31–40.
[2] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. *CoRR* abs/1705.02364 (2017). arXiv:1705.02364
[3] Hua He, Kevin Gimpel, and Jimmy Lin. 2015. Multi-perspective sentence similarity modeling with convolutional neural networks. In *EMNLP*. 1576–1586.
[4] Nitin Jindal and Bing Liu. 2006. Identifying Comparative Sentences in Text Documents. In *SIGIR (SIGIR '06)*. ACM, New York, NY, USA, 244–251.
[5] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *ICML*. 957–966.
[6] Xiang Ren, Yuanhua Lv, Kuansan Wang, and Jiawei Han. 2017. Comparative Document Analysis for Large Text Corpora. In *WSDM*. ACM, NY, USA, 325–334.

[3]SemEval STS Task Dataset