Using Conditional Probability for Discovering Semantic Relationships between Named Entities in Cultural Heritage Data

Jordan Stoikov

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences,
Acad. Georgi Bonchev Str., Block 8, Sofia, Bulgaria
jstoikov@shieldui.com

Abstract. This paper introduces a method for extracting information from various cultural heritage data sources using word embedding with feature extraction from the sentence and structured representation of the sentence. The focus is on discovering the named entities' relations including part of speech tags and positons tags by the means of conditional probability.

Keywords: Word Embedding, Conditional Probability, Cultural Heritage, Word Vector, Semantic Relationship.

1 Introduction

Cultural heritage (Culture: Definition of the cultural heritage, n.d.)is an important topic in the information society. Many cultural heritage institutions plan to collect cultural heritage artifacts in digital format in order to preserve cultural heritage.

Due to the extremely fast growing volumes of digital data in recent years, the scientific searches aimed at efficient digital data access and analysis are extremely relevant. This data explosion is a reality in Europe and worldwide and directly affects digitized cultural and historical content. The problem we face is providing the means to structure, process, manage and visualize this data. It is becoming clear that data is already an asset that can create a significant competitive advantage, but its structured digital form contributes to the correct extraction of knowledge from it.

The extraction of information is widely used for knowledge acquisition in the field of natural language processing (NLP). Knowledge is acquired from several data sources including structured data, semi-structured data, and unstructured data with the aim to determine named entities, relations and events, and then convert the extracted information to structured representations. Many approaches related to tasks of information extraction, focus on the determination of name entities and relationships between a pair of entities. The lack of semantic relationships in such approaches make it difficult to extract cultural heritage information into a meaningful relationship that would help to

Digital Presentation and Preservation of Cultural and Scientific Heritage. Conference Proceedings. Vol. 11, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS, 2021. ISSN: 1314-4006, eISSN: 2535-0366

determine the concept of semantic relation between entities. Therefore, these works may not be applied effectively in a number of applications such as question answering, information retrieval, and named entity recognition systems.

Recently, the idea of word embedding based on vector space is gaining popularity. In this approach the word embedding learns from the distributional information of words in corpora (a large structured set of texts). This method relies on distance or angle between word vectors for estimating the quality of a set of word representations. Word embedding can be utilized to capture semantic similarities, and can also be used to predict analogies between words.

This paper introduces methods and approaches for extraction of name entities and the triple semantic relationship between them from various sources of data. The model under review identifies named entities by using word embedding with feature extraction such as part of speech and position tags, using conditional probability to discover the probability relationship and create a knowledge graph, based on the extracted information (Paneva-Marinova, Stoikov, Pavlova , & Nikolova , 2020). The proposed approach works with accuracy for specific domains where similar arguments are associated with similar relations, resulting in the identification of the named entities and their sematic concept relations.

2 Background Knowledge and Related Work

Extraction of information (IE) is an essential task in the domain of cultural heritage. Significant information from various sources can be obtained, therefore making it possible to preserve cultural heritage related information for future generations. Information extraction consists of two main sub-tasks, the extraction of named entities (NER) such as person (PER), location (LOC) and (ORG) organization and the extraction of relations (RE) between them (Xia & Liu, 2016). In Figure 1 (Adnan & Akbar, 2019) is shown the architecture of a system for information extraction. The first step is collection of data, which implies the process of process of integrating raw text from several data sources to the repository. In the second step the raw information of the document is split into sentences by using segmenting the sentence, and then each sentence is subdivided into words using a tokenizer. In the third step, each sentence is tagged with part-of-speech tags, which is a form of word feature tagging. This step facilitates the searching the search for entities in each sentence. The last step uses relation detection in order to investigate possible relations between entities, resulting in the list of entities and relationships (Rong, 2016).

3 Word Vector Representation

Word2vec is a group of related models, using input as a large corpus of text and create vector space. These models are trained to reconstruct linguistic context of words. The primary word2vec models are as follows: Continuous Bag of Words (CBOW) (Soni, Shavlik , Shavlik, & Natarajan, 2016) and Skip-Gram. Figure 2 (Ling, Dyer, Black, &

Trancoso , 2015) shows the CBOW and Skip-Gram Model. It can be assumed that the current word in a sentence is w(t).

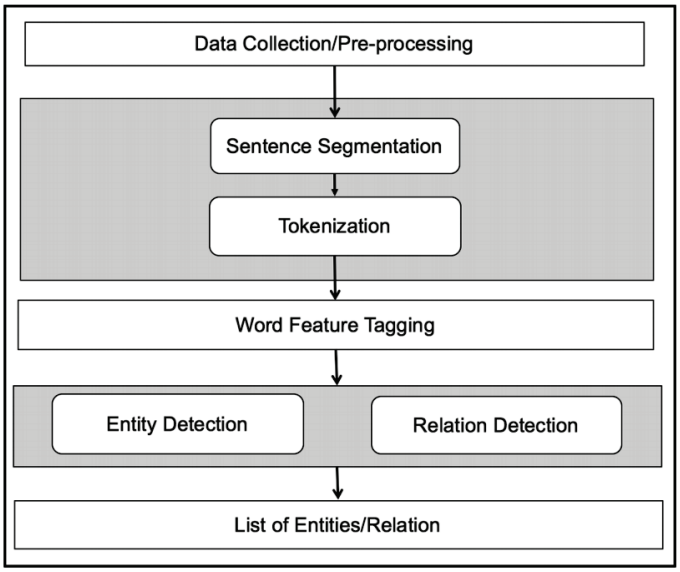

Fig. 1. Architecture of a system for information extraction.

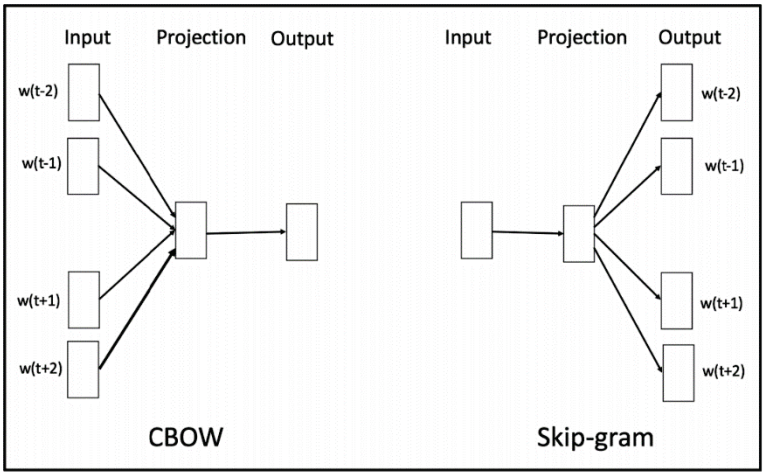

Fig. 2. The CBOW and Skip-Gram model.

CBOW: The task of this model is "predicting the word given a context". The model inputs are w(t − 2), w(t − 1), w(t + 1), w(t + 2). The model output will be w(t).

Skip-gram: This task of this model is the opposite. It predicts input word and the given context. The model input is w(t), and the outputs are w(t − 1), w(t − 2), wt + 1, w(t + 2).

4 Related Works

Word embedding techniques are used for embedding words into dimension vectors, based on the co-occurrence information of words and their context. Based on the short distances in the vector space is determined the similarities of the words. Syntactic and semantic similarities between words can be captured in these distances. The event extraction systems use word embedding vectors to represent language and address the problem of the limited labeled corpus by using a large amount of unlabeled data. Most of the relation extraction approaches primarily represent a pair of entities, and using dependency parsing relation extraction can solve the problem of identifying the patter of similarity between the relations.

The problem of dependency relation extraction is that the labeled data is expensive and limited. Another issue is classifiers trained with the limited data, resulting in findings that are biased with that corpus, leading to poor performance. In general, the models are limited to the syntactic and lexical patterns text. The results might lack some semantic relations. This knowledge is not enough when some information is missing or when the text is underspecified.

Recently, knowledge graph embedding has become attractive. This method embeds entities and relations in a knowledge graph into a vector space and manipulates this representation in a way that preserving the structure of the original graph. Hegde and Talukdar (Hedge & Taludar , 2015) present their Entity-Centric Expansion (ENTICE), an entity-centric framework for increasing knowledge densities in automatically constructed knowledge graphs. It is capable of extracting facts belonging to four types: known target entity, new target entity, known relation, and new relationships. Most of the existing knowledge graph embedding model can extract the individual triples. However, there is missing information because they lack the relationship of the entities connected to the same node, which relates to the other. Therefore, the solutions should identify named entities and determine the probability semantic concept relations between entities. As a result, the knowledge graph can be created and connected to the other related node.

5 System Overview

Various data sources contain data with different metadata standards and formats. Figure 3 (Adnan & Akbar, 2019) outlines the architecture of a system for information extraction. The data collected from the different data sources is accumulated in a central repository, from where information is extracted in a five steps approach: data collection, pre-processing, pre-trained embedding, feature extraction and semantic relation extraction.

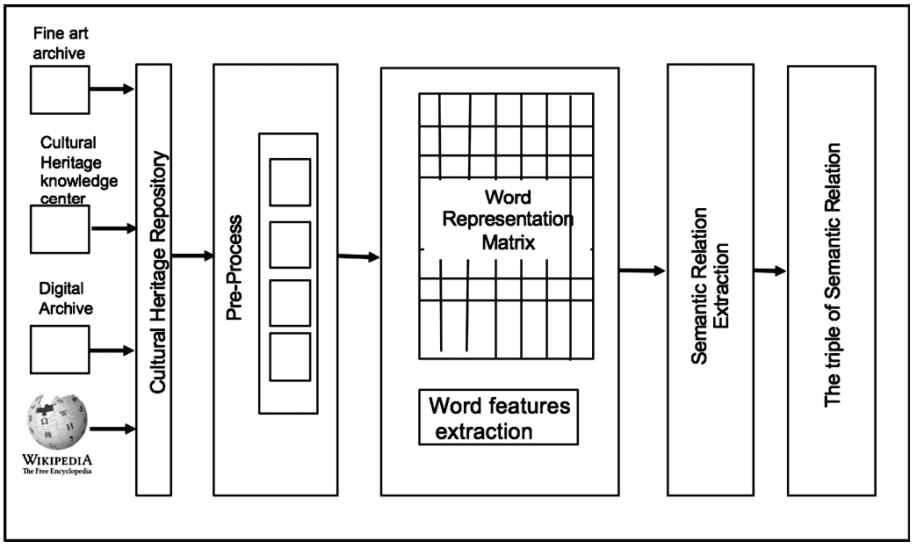

Fig. 3. Architecture of a system for information extraction.

5.1 Data Collection and Pre-process

The data is collected from cultural heritage data sources: (1) National cultural knowledge data centers (2) Wikipedia articles, focusing on cultural heritage domain.

Pre-training Word Embedding. The word embedding is a representation of a word. The word2vec (Milkov, Zweig, & Yih, 2013) was introduced as a model for a vector representation. The Skip-gram model is used to pre-train the word embedding. The Skip-gram model trains the embeddings of words w(1), w(2) … w(t) by maximizing the average log probability shown in Eq. (1)

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p(w_{t+j}|w_t) \tag{1}$$

where c is the window or context surrounding the current word being trained on. The basic Skip-gram formulation is to normalize this vector by using the soft-max function addressed in Eq. (2).

$$\mathrm{p}(w_O|w_I) = \frac{exp\left(v'_{wO}Tv_{wI}\right)}{\sum_{w=1}^{W}\exp\left(v'_{w}Tv_{wI}\right)} \tag{2}$$

5.2 Features Extraction

The set of features extracted from the sentence and structured representation of the sentence is used in the below-described approach.

Part of Speech (POS) Tagging. The process that classifies words into their parts of speech and labels them accordingly is known as part-of-speech tagging or POS tagging. Parts of speech are also known as lexical categories or word classes. The collection of tags that are used for a particular task is known as a tag set. This feature uses a description of five parts-of-speech: noun, verb, pronoun, preposition, and adverb. Figure 4 shows POS tagging and how classifier combination techniques can support the POS taggers.

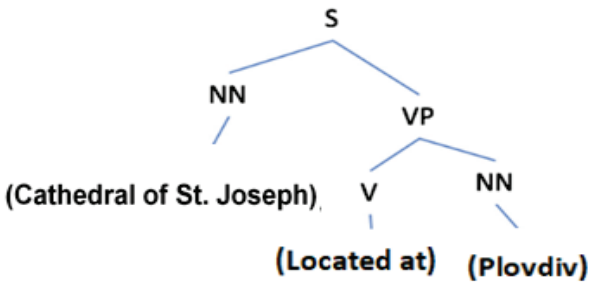

Fig.4. The part of speech tagging.

5.3 Dependency Relation Extraction

Dependency relations ae considered at the word level. Generally, the sentence structure follows the order of Subject-Verb-Object. In some cases, subjects and objects are excluded from the sentences. Therefore, some sentences consist only of a verb or a verb and an object. In our case study are utilized the part of speech and position features as follows: POS of the subject word and the object word of the sentences distance between both words, the dependency relations direction and relative position of the subject word and the object word. The basis binary tree is created based on the grammatical relation that contains the dependency structures. The best dependency structures found by the given prospective root position and the dependency matrix based on Nivre's Arc-Standard (Nivre, 2004). Figure 5 shows the example of dependency structure for a sentence.

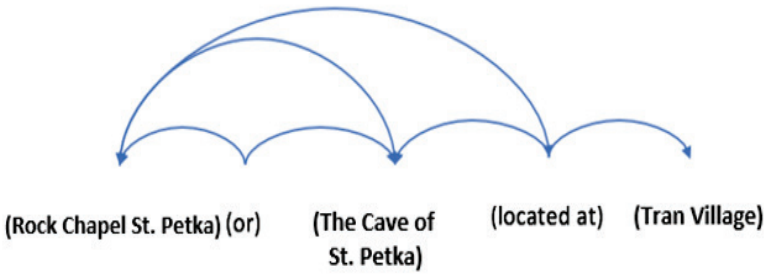

Fig. 5. Dependency structure for a sentence.

5.4 Semantic Relation Extraction

To conduct the relational similarity prediction is to assess the degree to a word pair (A, B) in the same relation as another pair (C, D), or to complete an analogy A: B :: C: ?. The analogy can be completed by:

$$d = \underset{x}{\operatorname{argmax}} \frac{(w_b - w_a + w_c)^T w_x}{\|w_b - w_a + w_c\|} \tag{3}$$

After an analogy is completed, is executed the following wb − wa = wd − wc (For instance, queen − king = actress − actor). The word vectors trained by the word2vec model can be used to effectively measure semantic similarity. In the relational formula wb − wa = wd − wc, for example, the country-capital city relationship: vec("Sofia") − vec("Bulgaria") ≈ vec("Athens") − vec("Greece"). Certain characteristics are applied in order to develop a similarity measure between two entities of analogy relation.

Table 1. The examples of types of analogies.

Types	Examples
Geography	Sofia: Bulgaria
Synonym	Rock Chapel St.Petka : The Cave of St.Petka

For example in the pair, (Cathedral of Saint Joseph: Sofia) and (Cathedral of Saint Louis: Plovdiv), is computed the embedding offsets between entities of each pair as follow: vec("Cathedral of Saint Joseph ") − vec("Sofia") ≈ vec("Cathedral of Saint Louis ") − vec("Plovdiv"). After that is calculated the cosine distance between their embedding offsets. Table 2 shows the similarities between two instances of the relations.

Table 2. Similarities between two instances of the relations.

Two entities	Similarity by cosine distance

(The Cave of St.Petka, Tran Village) & (Cathedral of Saint Louis , Plovdiv)	≈0.11
(Alexander Nevsky Cathedral, The St. Alexander Nevsky Cathedral) & (Rock Chapel St.Petka, The Cave of St.Petka)	≈0.08

In this paper, the focus is on a triple relations consisting of a subject, a predicate, and an object. The characteristics of the relational similarity prediction are applied to develop a similarity measure between two entities and focus on the predicate using POS. We are given a new triple of the form: A:?:B :: C:?:D For example, (The Cave of St. Petka: Located at: Tran Village) and (Church of St. Constantine and Helena: Located at: Plovdiv), we compute the embedding offsets between entities and predicate of each pair as follow:

$$\overrightarrow{A1} - \overrightarrow{B1} + \overrightarrow{A2} \approx \overrightarrow{B2} \tag{1}$$

$$\overrightarrow{A1} - \overrightarrow{B1} \approx \overrightarrow{A2} - \overrightarrow{B2} \tag{2}$$

$$\overrightarrow{A1} - \vec{P} \approx \overrightarrow{B1} - \vec{P} \text{ and } \vec{P} - \overrightarrow{A2} \approx \vec{P} - \overrightarrow{B2} \text{ Where } \mathrm{Max}(P(P|A1)) \tag{3}$$

We calculate the cosine distance between their embedding offsets. Table 3 shows the similarities between instances of the triple relation.

Table 3. The similarities between instances of triple relation.

Two entities with relation	Similarity by cosine distance	P(A|B) (Conditional probability)
(The Cave of St.Petka, Located at, Tran Village) & (Cathedral of Saint Louis , Located at, Plovdiv)	≈0.11	0.24
(Alexander Nevsky Cathedral, has name, The St. Alexander Nevsky Cathedral) & (The Cave of St.Petka, has name, Rock Chapel St. Petka)	≈0.08	0.20

6 Some Experiments

In this study were used datasets from various sources with focus on the domains: places, events, person, and artifact. One of the datasets is a Wikipedia article, focusing on cultural heritage domain. The articles represent the triples (for instance located at, built by) in each article. We follow pre-process step and divide the sentence into meaningful units. Tokenization is carried out with world lists built by dictionaries. Further is used

the Skip-Gram model and the softmax function in order to train two sets of word embedding. The resulting relation extraction is highly effective as compared with CIDCOM (CRM), a tool that provides extensible ontology for concepts and information in cultural heritage and museum documentation. It is the international standard (ISO 21127:2014) for the controlled exchange of cultural heritage information (Simov, 2019). The accuracy of relation extraction was determined after conducting test of the relation extraction task, using different feature sets and comparing the obtained results.

Table 4. Accuracy of relation extraction.

Model	Feature Set	Accuracy (%)
Word2vec (CBOW)	POS	40
Word2vec (Skip-gram)	POS	59
Word2vec (CBOW)	POS + position	70
Word2vec (Skip-gram)	POS + position	81

The similarities between instances of triple relation calculated in this study are used to generate the knowledge graph shown in Figure 6.

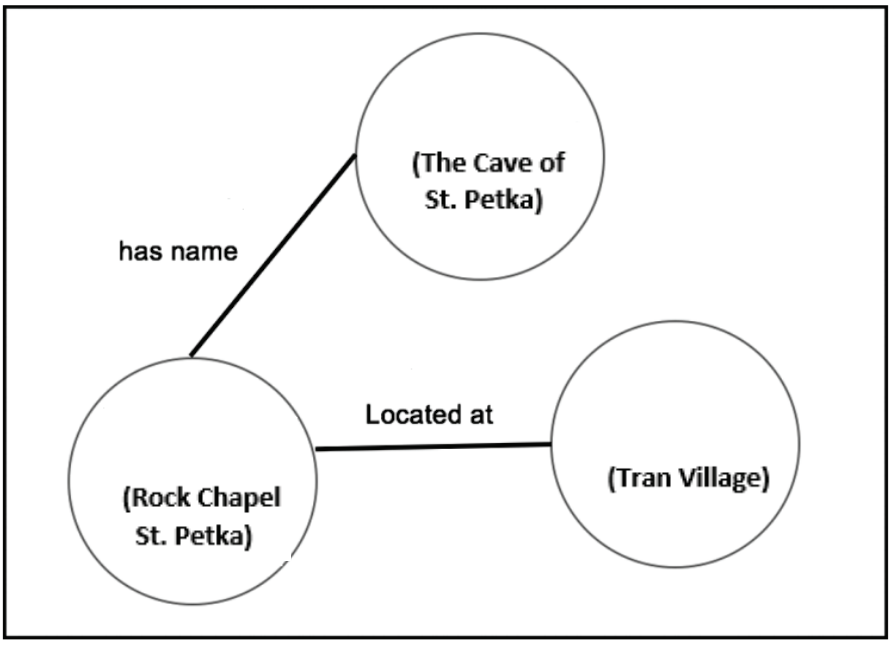

Fig. 6. Result for knowledge graph.

7 Conclusion

This paper introduces an approach to extracting information from various cultural heritage data sources. The focus is on identifying named entities and determining the semantic concept relation between entities. By utilizing word embedding with feature extraction, can be discovered entities' relations including part of speech tags and position tags. The method is evaluated using CIDOC(CRM), a common standard ontology for cultural heritage information. The results demonstrate high accuracy at 81%, which make this approach applicable in real-life data extraction from diverse data environments.

Acknowledgements

This work is funded in part by CLaDA BG, the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH, Grant number DO01-377/18.12.2020 and by the Bulgarian Science Fund under the research project №КР-06-N50/4 30.11.2020 "Fourteenth Century South Slavonic scribes and scriptoria (palaeographical attribution and online repertorium)" (https://kopisti14.kmnc.bg/bg/).

References

Adnan, K., & Akbar, R. (2019, October 17). An analytical study of information extraction from unstructured and multidimensional big data. *Journal of Big Data 6*, 7-10.

Culture: Definition of the cultural heritage. (n.d.). Retrieved 06 01, 2021, from https://en.unesco.org/

Hedge, M., & Taludar , P. (2015). An Entity-centric Approach for Overcoming Knowledge Graph Sparsity. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 530–535). Lisbon: Association for Computational Linguistics.

Ling, W., Dyer, C., Black, A. W., & Trancoso , I. (2015). Two/Too Simple Adaptations of Word2Vec for Syntax Problems. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1299–1304). Denver: Association for Computational Linguistics.

Milkov, T., Zweig, G., & Yih, W.-t. (2013). Linguistic Regularities in Continuous Space Word Representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746–751). Atlanta: Association for Computational Linguistics.

Nivre, J. (2004). Incrementality in Deterministic Dependency Parsing. *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together* (pp. 50-57). Barcelona: Association for Computational Linguistics.

Paneva-Marinova, D., Stoikov, J., Pavlova , L., & Nikolova , A. (2020). Data Discovery and Distributed Representation for Better Cultural Heritage Observation and Learning. *Proceedings of 13th annual International Conference of Education, Research and Innovation* (pp. 5696-5699). IATED: ISBN: 978-84-09-24232-0,ISSN: 2340-1095. Retrieved from https://library.iated.org/view/PANEVAMARINOVA2020DAT

Rong, X. (2016). *word2vec Parameter Explained*. Retrieved from https://arxiv.org/abs/1411.2738

Simov, K. (2019). Integrated Language and Knowledge Resources for a Bulgarian-Centric Knowledge Graph. *Digital Presentation and Preservation of Cultural and Scientific Heritage. Vol. 9* (pp. 65-74). Sofia: ISSN 1314-4006 (Print), eISSN 2535-0366 (Online).

Soni, A., Shavlik , V., Shavlik, J., & Natarajan, S. (2016). Learning Relational Dependency Networks For Relation Extraction. *Inductive Logic Programming* (pp. 81-93). New York: Springer.

Xia, Y., & Liu, Y. (2016). *Chinese Event Extraction Using DeepNeural Network with Word Embedding*. Retrieved from https://arxiv.org/abs/1610.00842

Received: June 15, 2021
Reviewed: July 05, 2021
Finally Accepted: July 15, 2021