Relation-Centric Semantic Annotation using Semantic Role Labeling and Coreference Resolution

Chia-Hung Lin[1], Sheng-Hao Hung[2], Chien-Hsiang Liao[3]
[1] Department of Management Science, R.O.C. Military Academy, Taiwan
[2] Computer and Network Center, National Chi Nan University, Taiwan
[3] Department of Information Management, National Central University, Taiwan

Abstract - *Automatic semantic annotation based on domain-specific ontologies is a one of the critical issues for the success of the semantic web. Most existing approaches focused on the detection of concepts such as named entities, dates, monetary amounts. This study explores automatic semantic annotation techniques for applications using relation-centric ontologies which represent domain knowledge using a set of concepts with many inter-class relations. We propose a framework to detect event-based concepts and inter-concept relations using semantic role labeling and coreference resolution techniques. We gave an illustration of the processes by a semantic annotation application using CIDOC-CRM as the underlying ontology. Experiments using archives with a large number of image descriptions were conducted. The primitive results show that the accuracy is about 80% or so.*

Keywords: Semantic Web, Semantic Role Labeling, Semantic Annotation, Ontology, Coreference Resolution.

1 Introduction

The development of Semantic Web is an important step to facilitate the knowledge-based information integration across different resources. To create document annotations with well-defined semantics, the Semantic Web proposes annotating document contents based on domain ontologies [3] that often formally identify "concepts" and inter-concept "relations" between concepts in a specific domain. One major challenge in the semantic annotation is that the annotations by human are often laborious and error-prone. In the Semantic Web community, there is a thirst for technologies that perform the semantic annotation in more automatic manners.

One approach for automating the semantic annotation from free-text resources is the "string-matching" technique. Based on the string-matching technique, in the literature, a number of tools have been developed for Semantic Web applications. For example, Popov et al. [28] and Kiryakov et al. [18] proposed frameworks for semantic annotation, indexing and retrieval. They applied name entity recognition technique to detect a variety of knowledge such as named entity, money amount in sentences. Other similar works that rely on string matching techniques for retrieving knowledge includes Open Ontology Forge, COHSE annotator, Mnm, Melita, Parmenides, Armadillo, SmartWeb, PANKow, KIM, and Magpie [2; 4; 5; 8; 9; 10; 11; 13; 27; 28; 33]. An extensive review of relevant studies on semantic annotation can be found in Uren et al. [32]. Overall speaking, as concluded in [32], most of these semantic annotation systems are designed mainly to recognize "concept" instances and values from texts, but they often are not able to establish explicit "relations" between concepts. Hence, the directly applicability of the existing string-matching approaches in applications using "relation-centric" ontologies, which represent domain knowledge using a set of concepts with many inter-class relations, is questionable. One such real-life relation-centric ontology is CIDOC-CRM (Conceptual Reference Model) which is designed for cultural heritage applications and the only one having acquired the status of an International Standard [7].

The primary role of the CIDOC-CRM is to enable information exchange and integration between heterogeneous sources of cultural heritage information. It aims at providing the semantic definitions and clarifications needed to transform disparate, localised information sources into a coherent global resource. The CIDOC CRM contains classes and logical groups of relations (properties). Those relation groups are used to express facts regarding identification, classification, participation, structure and parthood, location, influence and motivation, assessment and reference. The CRM can describe the semantics of hundreds or more schema in use for museum object documentation with a small set of 90 concepts and 148 properties.

In the CRM, the essential knowledge in cultural heritage domain is structured as a semantic net with classes and the associated properties between classes. Most of the relationships people intuitively describe between classes are actually deductions from specific kinds of "events." For example, an *E67 Birth Event* comprises several relations to relevant concepts, such as event participants, time, and location (Figure 1).

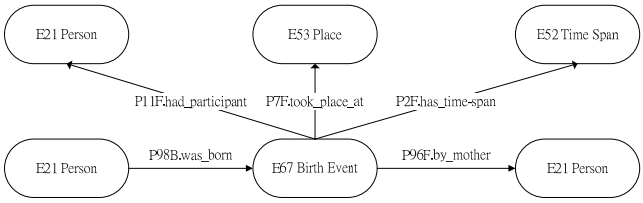

Figure 1: The CRM representation of a birth event

Considering the complexity of the CRM classes and relations, as in most Semantic Web applications, annotations by human based on the CRM are certainly laborious and error-prone. Current, there is a need for automatic semantic annotation techniques in cultural heritage communities. Unfortunately, existing string-matching based approaches fail to effectively identify inter-class relations that are required in CRM-based applications. The purpose of this study is to explore technologies that perform such a "relation-centric" semantic annotation in a more automatic manner. We propose mechanisms to retrieve event related knowledge from free-text resources using state-of-the-art natural language processing techniques. The objective is to automatically detect particular CRM-based event-related information, such as, subject, object, time and location in texts and map the knowledge into CRM instances for Semantic Web applications. In this following, the underlying principles and process for our methodology will be elaborated.

2 Methodology

In this paper, we propose a framework that adapt and integrate a number of state-of-the-art natural language techniques, including, lexical pattern-based knowledge retrieval, semantic role labeling, and coreference resolution techniques. In the first stage, a possible CRM event is obtained by matching the texts to a number of lexico-syntactic patterns that unambiguously represent the desired semantic relations. In the second stage, the core semantic roles, such as subject, object, location and temporal in the extracted sentences are obtained using semantic role labeling techniques. In the third stage, coreference techniques are applied to identify the exact entity the extracted subject or object of the sentence refers to. In the final stage, the coreference-resolved semantic roles are mapped into corresponding CRM instances based on a variety of mapping rules. In the following, we elaborate the underlying principles and the detailed operations of each stage.

2.1 Stage 1: sentence retrieval based on lexical-syntactic pattern matching

In principle, automatic discovery of a particular knowledge must start with a thorough investigation of the lexical terms and syntactic forms used to reliably express the desired semantic relation between entities. In practice, there are varieties of lexico-syntactic patterns that express a particular desired CRM event (e.g., creation, modification, destruction) describing a cultural artifact. To accumulate such syntactic patterns, we first come out with certain popular "seed events" which are expected to be widely described in abundance of web pages. Based on the seed events, appropriate query strings are formulated to query web search engines. The retrieved snippets from the web search results are investigated manually to identify the syntactic patterns that unambiguously indicated the desired event.

We take the "donation" event as an example to illustrate the syntactic pattern collection process. We first come out with a well-known event, for example, "France donates Statue of Liberty to U.S", as the seed events to query the web search engines. Examples of sentences retrieved that are used to explicitly refer to the knowledge come from the web search results at least include:

- France *donated* the Statue of Liberty to the USA
- It was in recognition of this that France *bequeathed* the Statue of Liberty to New York City in 1886.
- During the centennial France *offered* the Statue of Liberty as a gift.
- The people of France *presented* the Statue of Liberty to the minister of the United States in Paris.
- France *sent* the Statue of Liberty as a gift to the U.S. in order to celebrate.
- The people of France *gave* the Statue of Liberty to the people of the United States in 1886 in recognition of the friendship established during the American Revolution …

Based on the collected sentences, we realize a donation event possibly exist if a sentence contains phrases such as present, donate, bequeathed, "offered … as a gift", "send … as a gift", gave, etc. Raw sentences containing these listed terms will be candidates that possibly contain the desired event. These candidate sentences are collected and feed into semantic role labeling and coreference processor for distilling the knowledge desired. To increase the coverage of the query results, the query formulations are expanded by incorporating morphological variations such as verb tenses.

2.2 Stage 2: semantic role identification using semantic role labeling

Once the candidate sentences that contains a desired event that are included in CRM, the next stage is to identify the subject, object, location, and temporal

information associated with the event. We use the semantic role labeling technique to achieve this goal. A general overview on the state-of-the-art semantic role labeling techniques can be found in [6; 14; 25]. Roughly speaking, in a sentence, a verb (predicate) indicates an event. The verb's syntactic arguments generally are associated with the participants of the event. A semantic role is the relationship that a syntactic argument has with the verb. One of the most commonly-used schemes for specifying the semantic roles are proposed to construct a large-scale corpus - the PropBank [17; 21]. In PropBank, the arguments of a verb are labeled sequentially from ARG0 to ARG5, where ARG0 is usually the subject of a transitive verb; ARG1, its direct object, etc. A variety of adjunctive arguments, such as ARGM-LOC, for locative, and ARGM-TMP, for temporal, are also tagged. Semantic role labeling techniques automatically identify the semantic roles of a sentence. Automatically tagging the semantic roles with high precisions is difficult since an event can often be referred using varieties of lexical items with different syntactic realizations. In the literature, there are a number of studies proposed different methodologies for such purpose, i.e., [14; 19; 24]. These methodologies have obtained well accurate results about 88% on ARG0, 82% on ARG1, and 70% on ARGM-LOC, ARGM-TMP, for sample data from Wall Street Journal [19].

In this study, the SRL technique is applied to obtain this fine-grained information associate with the event. As an example, consider a sentence in the description texts for the artifact "Tomb of Pope Paul III", give as:

In 1628 "Tomb of Pope Paul III" was modified by Bernini.

The SRL results for the given sentence is given as

[$_{ARGM\text{-}TMP}$ In 1628] [$_{ARG1}$ Tomb of Pope Paul III] [$_{Target}$ was modified] [$_{ARG0}$ by Bernini]

In such a case, the obtained semantic roles are actually ready for mapping to CRM instances. Nevertheless, as described below, there are many situations that coreference resolution techniques are required to give a serviceable annotation.

2.3 Stage 3: Coreference Resolution for semantic roles

In the free-text descriptions for artifacts in cultural archives, quite often a complete semantic relation is expressed in different contextual sentences. The coreference needs to be resolved automatically to identity which entity a noun phrase or pronoun actually refers to. For example, in a sentence give as:

In 1628 "Tomb of Pope Paul III" was modified by him.

The ARG0 in this case is given by a SRL tool as "him", which surely does not give a valuable knowledge when mapped to a CRM instance. In such case, the coreference techniques need to be applied to resolve the coreference so as to identify the exact roles implied in neighboring sentences. In linguistics, coreference occurs when different expressions in a sentence or contextual sentences refer to a same entity in real world. Two expressions (noun phrases or pronouns) are said to be co-referring to each other if both of them resolve to a unique entity (i.e., the referent) unambiguously. For example, in the sentences, "Leonardo da Vinci was one of the greatest painters of the Italian Renaissance. He left only a handful of completed paintings, among his works, the Mona Lisa is the most famous painting", the "Leonardo da Vinci" and "he" are most likely coreferent. Coreference resolution is the task of resolving noun phrases or pronouns to the entities that they refer to. It has been an active research topic in natural language processing for decades. The coreference resolution techniques are widely used in areas such as named entity extraction, question answering, machine translation and so on. In the literature, quite a number of methodologies have been proposed for solving the coreference resolution. Most early attempts heavily rely on linguistic and domain knowledge [15]. On the other hand, most recent approaches apply machine learning techniques with sophisticated syntactic parser and tagger, e.g., [16; 23; 26].

2.4 Stage 4: Mapping semantic roles to CRM instances

Once the semantic roles of a candidate sentence are extracted with coreference resolved, certain event-specific heuristic rules will be applied so as to correctly map the semantic roles into a sensible CRM instances. The rules are manually designed based on extensive investigations on the possible syntactic constituents of the sentences containing the event-trigger patterns. For example, in many sentences, certain heuristic rules are required to filter undesired relevant information about the person. For example, in the sentence "Fra Angelico, a famous painter, was born in Guido di Pietro", the parsed ARG0 is given as "Fra Angelico, a famous painter". In such a case, only the proper noun "Fra Angelico" is mapped to the CRM instance. The detections of proper noun can often be done using coreference tools.

3 Primitive Evaluations

We carried out an experiment to investigate the performances of the proposed methodology in real life cultural digital archives. First, a large set of images with textual descriptions are collected from a number of online archives, including Louvre Museum[20], Web Gallery of ART[34], Rijksmuseum[29], Manchester Art Gallery [22] and The Metropolitan Museum of ART[30].

A collection of 30,300 artifacts and 173,000 sentences were taken from the five archives. The average sentence numbers in a painting description is about 10. The average word number in a sentence is 22. The textual data are parsed, sentence-by-sentence, using a public available semantic role labeling engine- ASSERT [24]. The parsed semantic roles for each image are managed in a database. For the coreference resolution, we applid the "Gate tool" [12]. The approaches of Gate can be found in http://gate.ac.uk/. A list of the artist names given in ULAN [31] are feed to the Gate tool such that the proper nouns of persons can be successfully detected. A variety of heuristic rules have been manually designed so as to map the semantic roles in a sentence to a corresponding CRM event instances.

Table 1: A list of lexico-syntactic patterns used in the experiments

CRM Event	Lexico-syntactic patterns
E6 Destruction	was destroyed
	ruin
	caused damage to
E8 Acquisition	bequeath
	give
	offer
	send
	made a contribution of
	endow
	contribute
	donate
	made a donation of
E11 Modification	adapt
	modify
	make alteration in
	made an amendment to

For the evaluation experiment, we applied the proposed approaches to detect 3 core CRM events, including E6 Destruction, E8 Acquisition and E11 Modification. Table 1 lists the lexico-syntactic patterns that are used to access the target raw sentences. Table 2 lists the evaluation results. We measured the retrieval effectiveness by precision rate [1]. Precision is the number of relevant items retrieved as percentage of the total number of items retrieved. Precision will degrade by incorrectly relevant items. Hence, Precision is mainly to measure the ability of a system to present only relevant items.

$$precision = \frac{Number\ of\ relevant\ items\ retrieved}{Total\ number\ of\ items\ retrieved}$$

Table 2 shows the corresponding precision rate for three CRM events. The precision rates, ranging from 79 to 83 %, appear to be fair satisfactory. Based on the observation on those instances that are un-correctly mapped, the major source of errors was originated from the erroneous parsing result from the semantic role labeling tool applied. In the future, with the possible improvement on the state-of-the-art semantic role technologies, the proposed approach in the paper appears to be promising to get higher precision results.

Table 2: Evaluation results

CIDOC CRM Event	Precision
E6 Destruction Event	83%
E8 Acquisition Event	80%
E11 Modification Event	79%

4 Conclusions

In this paper, we proposed a methodology for automatically retrieving event-based knowledge for semantic annotation from texts. We applied state-of-the-art natural language techniques, including the semantic role labeling and coreference techniques to achieve the goal. We use a well-developed relation-centric ontology in cultural domain – CIDOC CRM to illustrate the semantic annotation process. The evaluation results show that the accuracy is rather satisfactory. The ease of implementation also indicates that the proposed methodologies can be easily realized using public-available resources.

5 Acknowledgement

This work was supported by the National Science Council of Taiwan (R.O.C.), under grant NSC 99-2410-H-145-002.

6 References

[1] Baeza-Yates, R. and Ribeiro-Neto, B. "Modern Information Retrieval." Addison-Wesley Longman Publication Co., Inc., Boston, MA, pp. 24-60, 1999.

[2] Bechhofer, S., Goble, C., Carr, L., Hall, W., Kampa, S. and Roure, D.D. "COHSE: Conceptual Open Hypermedia Service." In S. Handschuh, S. Staab (Eds.), Annotation for the Semantic Web, IOS Press, Amsterdam, 2003.

[3] Berners-Lee, T., Hendler, J. and Lassila, O. "The Semantic Web." Scientific American, 2001.

[4] Black, W.J., McNaught, J., Vasilakopoulos, A., Zervanou, K., Theodoulidis, B. and Rinaldi, F. "CAFETIERE Conceptual Annotations for Facts, Events, Terms, Individual Entities, and Relations." Parmenides Technical Report, TR-U4.3.1, 2005.

[5] Buitelaar, P. and Ramaka, S. "Unsupervised Ontology Based Semantic Tagging for Knowledge

Markup." In Proceedings of the Workshop on Learning in Web Search at the International Conference on Machine Learning, Bonn, Germany, 2005.

[6] Carreras, X. and Màrquez, L. "Introduction to the CoNLL-2005 shared task: Semantic Role Labeling." In Proceedings of the CoNLL-2005, pp. 152-164, 2005.

[7] CIDOC CRM, available at: http://cidoc.ics.forth.gr/, last visited Apr. 30, 2011.

[8] Cimiano, P., Handschuh, S. and Staab, S. "Towards the Self-Annotating Web." In Proceedings of the 13th International World Wide Web Conference (WWW 2004), New York, NY, 2004.

[9] Ciravegna, F., Chapman, S., Dingli, A. and Wilks, Y. "Learning to Harvest Information for the Semantic Web." In Proceedings of the First European Semantic Web Symposium, Heraklion, Greece, 2004.

[10] Ciravegna, F., Dingli, A., Petrelli, D. and Wilks, Y. "User-System Cooperation in Document Annotation Based on Information Extraction." In Proceedings of the 13th International Conference on Knowledge Engineering and KM (EKAW02), Spain, 2002.

[11] Collier, N., Kawazoe, A., Kitamoto, A., Wattarujeekrit, T., Mizuta, Y. and Mullen, A. "Integrating Deep and Shallow Semantic Structures in Open Ontology Forge." In Proceedings of the Special Interest Group on Semantic Web and Ontology, JSAI (Japanese Society for Artificial Intelligence), Vol. SIG-SWO-A402-05, 2004.

[12] Dimitrov, M., Bontcheva, K., Cunningham, H. and Maynard, D. "A Light-weight Approach to Coreference Resolution for Named Entities in Text." In Proceedings of the Fourth Discourse Anaphora and Anaphor Resolution Colloquium (DAARC), Lisbon, 2002.

[13] Dzbor, M., Motta, E. and Domingue, J. "Opening up Magpie via Semantic Services." In Proceedings of the 3rd International Semantic Web Conference, Hiroshima, Japan, 2004.

[14] Gildea, D. and Jurafsky, D. "Automatic Labeling of Semantic Roles." Computational Linguistics, Vol. 28, No. 3, pp. 245-288, 2002.

[15] Hobbs, J.R. "Resolving pronoun references." In Barbara J. Grosz, Karen Sparck-Jones, and Bonnie L. Webber, editors, Readings in Natural Language Processing, pp. 339-352, Morgan Kaufmann, Los Altos, California, 1986.

[16] Kehler, A., Appelt, D., Taylor, L. and Simma, A. "The (Non)utility of Predicate-argument Frequencies for Pronoun Interpretation." In Proceeding of 2004 North American chapter of the Association for Computational Linguistics annual meeting, pp. 289-296, 2004.

[17] Kingsbury, P. and Palmer, M. "From Treebank to PropBank." In Proceedings of the LREC, 2002.

[18] Kiryakov, A., Popov, B., Terziev, I., Manov, D. and Ognyanoff, D. "Semantic Annotation, Indexing, and Retrieval." Journal of Web Semantics, Vol. 2, No. 1, 2005.

[19] Koomen, P., Punyakanok, V., Roth, D. and Yih, W-T. "Generalized Inference with Multiple Semantic Role Labeling Systems." In Proceedings of the Ninth Conference on Computational Natural Language Learning, CoNLL-2005, pp. 181-184, 2005.

[20] Louvre Museum. available at: http://www.louvre.fr/, last visited Apr. 30, 2011.

[21] Marcus, M., Kim, G., Marcinkiewicz, M.A., MacIntyre, R., Bies, A. Ferguson, M., Katz, K. and Schasberger, B. "The Penn-treebank: Annotating Predicate Argument Structure." In ARPA Human Language Technology Workshop, 1994.

[22] Manchester Art Gallery. available at: http://www.manchestergalleries.org/, last visited Apr. 30, 2011.

[23] Ng, V. and Cardie, C. "Improving Machine Learning Approaches to Coreference Resolution." In Fortieth Anniversary Meeting of the Association for Computational Linguistics, ACL-02, pp. 104-111, 2002.

[24] Pradhan, S., Ward, W., Hacioglu, K., Martin, J. and Jurafsky, D. "Shallow Semantic Parsing Using Support Vector Machines." In Proceedings of the Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting, pp. 233-240, 2004.

[25] Pradhan, S., Hacioglu, K., Ward, W., Martin, J. and Jurafsky, D. "Semantic Role Parsing: Adding Semantic Structure to Unstructured Text." In Proceedings of the International Conference on Data Mining (ICDM 2003), 2003.

[26] Ponzetto, S.P. and Strube, M. "Semantic Role Labeling for Coreference Resolution." In Companion Volume of the Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics, pp. 143-146, 2006.

[27] Popov, B., Kirayakov, A., Ognyanoff, D., Manov, D. and Kirilov, A. "KIM—A Semantic Platform for Information Extraction and Retrieval." Natural Language Engineering, Vol. 10, No. 3-4, pp. 375-392, 2004.

[28] Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., Kirilov, A. and Goranov, M. "Towards Semantic Web Information Extraction." In Proceedings of the Human Language Technologies Workshop at 2nd International

Semantic Web Conference (ISWC2003), Florida, USA, 2003.

[29] Rijksmuseum. available at: http://www.rijksmuseum.nl/ , last visited Apr. 30, 2011.

[30] The Metropolitan Museum of Art. available at: http://www.metmuseum/.org, last visited Apr. 30, 2011.

[31] ULAN, Union List of Artist Names, available at: http://www.getty.edu/research/conducting_research/ vocabularies/ulan, last visited Apr. 30, 2011.

[32] Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E. and Ciravegna, F. "Semantic Annotation for Knowledge Management: Requirements and a Survey of the State of the Art." Journal of Web Semantics: Science Services and Agents on the World Wide Web, pp. 14-28, 2005.

[33] Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A. and Ciravegna, F. "MnM: A Tool for Automatic Support on Semantic Markup." KMi Technical Report, TR No. 133, 2003.

[34] Web Gallery of Art. available at: http://www.wga.hu/, last visited Apr. 30, 2011.