Journal of Universal Computer Science, vol. 22, no. 6 (2016), 742-759
submitted: 30/11/15, accepted: 28/5/16, appeared: 1/6/16 © J.UCS

Automatic Generation of Interactive Cooking Video with Semantic Annotation

Kyeong-Jin Oh
(Department of Information Engineering, Inha University, Incheon, Korea
okjkillo@eslab.inha.ac.kr)

Myung-Duk Hong
(Department of Information Engineering, Inha University, Incheon, Korea
hmdgo@eslab.inha.ac.kr)

Ui-Nyoung Yoon
(Department of Information Engineering, Inha University, Incheon, Korea
entymos@hotmail.com)

Geun-Sik Jo
(Department of Information Engineering, Inha University, Incheon, Korea
gsjo@inha.ac.kr)

Abstract: Videos are one of the most frequently used forms of multimedia resources. People want to interact with videos to find a specific part or to obtain relevant information. To support user interactions, current videos should be transformed to interactive videos. This paper proposes an interactive cooking video system to generate automatically interactive cooking videos. To do this, the proposed system performs semantic video annotation on cooking videos. Semantic video annotation process includes three parts: synchronization between recipes and corresponding cooking videos based on a caption-recipe alignment algorithm, information extraction on food recipes using lexico-syntactic patterns, and semantic entity interconnection between recognized entities and semantic web entities. Cooking video annotation ontology is modeled to handle annotation data. To evaluate the proposed system, comparative experiments are performed on the caption-recipe alignment algorithm. The accuracy of information extraction and semantic entity interconnection is also measured. Experimental results show that the proposed system is superior to compared algorithms in alignment perspectives. Information extraction and semantic interconnection method also achieve high accuracy over 95%, respectively. Consequently, the proposed system generates interactive cooking videos in high accuracy and support user interactions by providing a user interface which allows users to easily find specific scenes and obtain detailed information on objects users have interested in.

Keywords: Interactive Cooking Video, Caption-Recipe Alignment, Entity Identification, Semantic Video Annotation, Cooking Video Annotation Ontology
Categories: L.3.0, L.3.2, M.0, M.1, M.7

1 Introduction

Everyday a huge volume of multimedia data are generated and consumed by people with the popularity of video sharing sites like YouTube [YouTube, 15], and Vimeo [Vimeo, 15]. The most widely consumed videos on the web can be roughly

categorized into two groups: entertainments and how-to videos. Entertainment videos contain video types such as music, comedy, talent, games, and etc. How-to videos as instructional videos guide people step by step on how to perform a specific task with clear demonstrations. While people normally just watch entertainment videos, they sometimes want to interact with videos to understand thoroughly each step in a case of watching how-to videos. The interactions with how-to videos can be divided into two types. First interaction is to control a seek bar on a video player. People often want to play videos from a specific part that they want to watch. In addition, they want to repeat a specific part because the part is difficult to understand for them. The second is an interaction with objects in videos. People have an interest in certain object appeared in a video during watching the video, and want to know the detailed information relevant to the object. They then have to search the information through search engines with keywords. However, sometimes they do not recall appropriate keywords for the object. Although they remember exact keywords, there are no guarantees that the results provide correct information because search engines just show a list of search results made by keyword-based indexing. Although video sharing services such as YouTube and Vimeo provide billions of video to users and manage them, they do not provide the interactive functions to user. To support user interactions within how-to videos, some annotation tasks on videos are required. Annotation data should be merged into videos and video players provide interaction functions with videos. In other words, current videos should be transformed to interactive videos.

This research focuses on cooking video domain among various multimedia domains. Cooking videos is a kind of how-to videos and many people refer cooking videos to cook their foods. People want to find a specific part of a cooking video and they want to know information on objects which is appeared in cooking video, such as ingredients, cooking tools and methods for ingredient preparation. However, current cooking video services such as All Recipes [AllRecipes, 15], BBC food [BBC Food, 15] and Foodista [Foodista, 15] do not provide these functions. It is not easy to quickly find a specific part that a user wants to watch because users also do not know that which part of a cooking video belongs to a certain instruction of a recipe. In addition, users cannot obtain information about objects appeared in a video due to a lack of annotation on cooking videos.

In this paper, an interactive cooking video system is proposed to solve the current issues on a cooking video domain. The proposed system automatically synchronizes each instruction of recipes to corresponding part of cooking videos through a caption-recipe alignment algorithm. To identify targets of user's second interaction abovementioned, ingredients and cooking tools information used in each instruction of a recipe are recognized by using lexico-syntactic patterns (LSPs) and semantically interconnects objects and semantic web entities. To manage annotation data semantically, ontology for interactive cooking videos is modeled and applied to the proposed system. The proposed system supports user interactions on interactive cooking videos generated by semantic cooking video annotation.

The rest of this paper is organized as follows. The next section describes related works. Section 3 presents the proposed interactive cooking video system based on semantic video annotation. Section 4 presents evaluation of the proposed system and a

user interface to play interactive cooking videos. Section 5 concludes the proposed approach and identifies future works.

2 Background Knowledge

2.1 Video Annotation

Videos become very popular multimedia resources in the Web since video sharing sites have emerged. Users want to search videos, find specific parts within videos and obtain information about object appeared in videos. Due to the rapid increase of the amount of videos, handling video data has become an important issue. To support user's behaviors on videos, many researchers focus on the handling of videos and address video annotation tasks plays an essential role in handling videos.

Image processing-based approaches apply machine learning techniques such as object recognition and object tracking for automatic video annotation [Ballan, 11] [Wu, 12]. Image processing-based video annotation techniques succeed to get high-level features by recognizing objects appeared in videos. However, their annotation performance depends on the accuracy of object recognition with domain characteristics. Accordingly, it is difficult to apply various domains. Another approaches for video annotation have applied text information [Cour, 08] [Ronfard, 03] [Turetsky, 04]. Textual metadata is considered very important factors to enable information extraction of video resources. However, because just few video resources provide textual information, a task to create textual metadata is required before the start of video annotation tasks. Although the textual data is offered from resource owners, metadata for video annotation should be extracted from the textual data. A metadata schema is also required to semantically and structurally manage annotation data for supporting video search, pin-point access, and user interaction with objects appeared in videos.

In this paper, we applied a text-based approach for generating interactive cooking videos. Closed captions of cooking videos and food recipes are used for the proposed approach.

2.2 Interactive Video

Interactive video is a video type that supports user interaction on a video. As new instructional technology, interactive video incorporates information management and decision-making capacity within video capabilities [Bellman, 09] [Homer, 14]. Interactive video not only play a video like a normal video file, but also contain clickable areas which is called "hotspots". Hotspots within an interactive video allow users to interact with the video when they click on them. User interactions on hotspots can be classified into two types. When a user clicks on a hotspot for an object, interactive video displays information about the object clicked by the user. If a hotspot is not for an object, interactive video moves to a different part of the video or open another video. How-to videos like cooking videos usually require these user interactions than other videos. A user may watch a how-to video from a specific part of the video or repeat the part without watching the video from begin to end. Those who make how-to videos sometimes cannot provide a detailed explanation for some

objects appeared in a video because if the explanation spends long time inadvertently, then the video may lost its focus. Therefore, if a how-to video is made in the form of an interactive video, both video producers and consumers can satisfy their needs through the interactive video.

Video annotation is a core technique for generating interactive videos. To annotate videos to make interactive videos, it is essential to acquire time for points to annotate and information to be offered at the time. However, it is difficult to automatically obtain the information.

Wirewax [Wirewax, 15] and Popcorn Maker [Popcorn Maker, 15] which are interactive video annotation tools support interactive video manufacturing in manual way. To make an interactive video, a user tries to find a part to annotate. The user then specifies a hotspot as an area of current frame for video annotation. Lastly, the user designates an action that either an interactive video displays the information about object the user clicks on or opens another video. Using these interactive video annotation tools, it is possible to make the definitive interactive videos, but the accountability for annotation tasks are delegated to users.

3 Automatic Generation of Interactive Cooking Video

In this section, system architecture and semantic video annotation process for the proposed interactive cooking video system is presented.

3.1 System Architecture

Fig. 1 presents system architecture for the proposed interactive cooking video system. At the resource crawling phase, a web crawler extracts cooking videos and corresponding food recipes from a cooking video web site. To extract closed captions of cooking videos, YouTube's auto caption feature is applied.

Alignment manager performs a caption-recipe alignment algorithm using food recipes and closed captions to identify the time information of cooking video parts corresponding to recipe instructions. Recipe information extractor extracts entities such as ingredients, cooking tools and ingredients portions from each recipe instruction. Extracted information is used to annotate concise information on hotspots in interactive cooking videos. Entity Interconnector interconnects extracted entities from the recipe information extractor with semantic entities having detailed information. Cooking Video Annotation (CVA) ontology is modeled to tackle food recipe and video domains. Semantic annotator generates interactive cooking videos by combining the results of alignment manager, recipe information extraction and entity interconnector with CVA ontology. All interactive cooking videos through semantic video annotation are saved. When a user watches an interactive cooking video, the proposed cooking video player displays the interactive cooking video and the user can interact with the video.

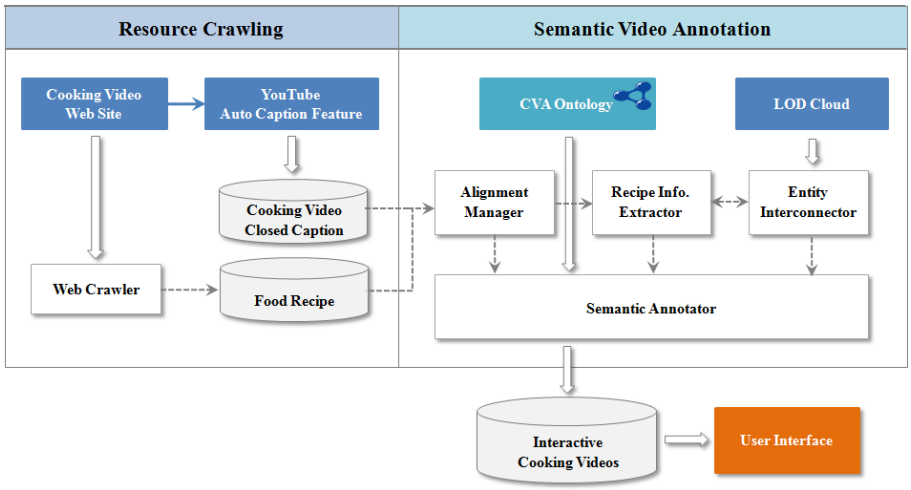

Figure 1: System architecture for the proposed interactive cooking video system

3.2 Cooking Video Annotation Ontology Modeling

An appropriate metadata schema is required to manage annotation data of cooking videos. In this paper, CVA ontology is modeled to manage annotation data in a linked data form. Fig. 2 shows some parts of the modeled CVA ontology used for the proposed system. The main purpose of CVA ontology is to manage recipe information with video information. Recipe-related classes such as Recipe, Tool, Recipe Step, and Ingredient Portion is from the linked recipe schema. In addition, CVA ontology extends recipe-related properties which are not modeled at the linked recipe schema, and includes properties to connect food recipes and cooking videos.

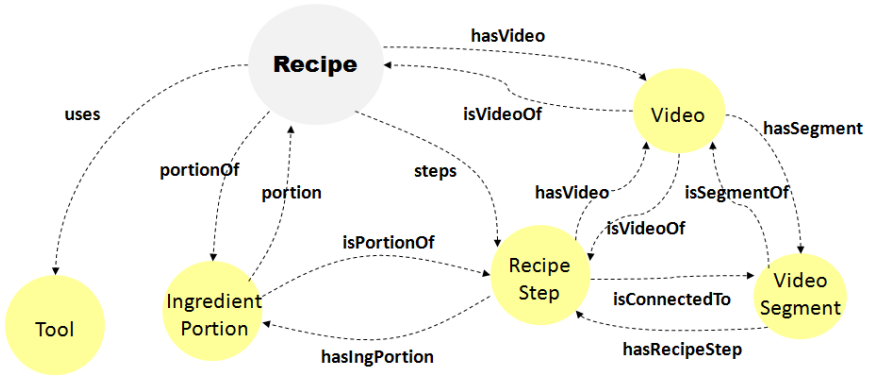

Figure 2: Cooking video indexing and annotation ontology

Some of additional properties are shown in Table 1.

Property	Description
nutrient	Each ingredient used in recipe has nutrient information
preparationTime	A time for preparation of ingredients and tools for cooking
cookingTime	A Time for real cooking process from the start of recipe steps
isVideoOf	A relation between Recipe and Video class
isConnecteTo	A relation between Recipe Step and Video Segment Class
hasSegment	A relation to specify that a video is segmented according to a step

Table 1: Properties of Cooking Video Ontology

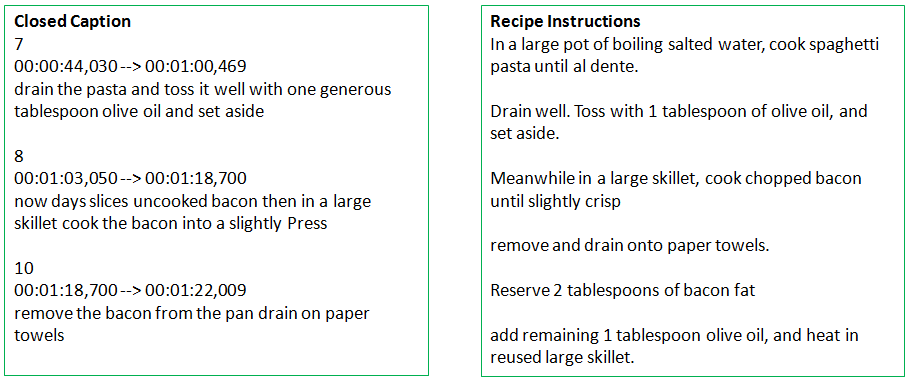

Figure 3: An example of closed caption and recipe

Caption-recipe alignment task is to find subsequence which seems to be virtually identical. Therefore the task is mapped to identification of LCS (Longest Common Subsequence) between two different documents. LCS problem is solved by applying DP (Dynamic Programming) technique and many researchers have applied DP to solve identification of LCS in their researches [Cour, 08] [Lambert, 13] [Ronfard, 03] [Turetsky, 04]. By using DP algorithm, they are able to find optimal path between given two sequences and detect an optimal alignment between script and subtitles. In contrast with the existing processes, the proposed caption-recipe alignment algorithm is based on sentence of recipe and text captions extracted by cooking videos. While chef explains every step of a recipe with demonstrations, recipe instructions is constructed in the form of imperative sentence. Also, sometimes, the sequence of recipe steps can be different in cooking video because some recipe steps are regardless of the sequence used. It does not matter to people, but can cause problem to a machine which handle the sentences. Accordingly, for a caption-recipe alignment process, a similarity matrix is based on sentences of both a text caption and a recipe and optimal path is found on the matrix.

As mentioned above, closed captions from the auto caption feature of YouTube is not perfect and the captions are divided into each element based on speech interval of

speakers in a cooking video. Alignment manager reconstruct the captions so that each element is based on a sentence.

Fig. 4 represents a caption-recipe alignment algorithm.

```
Input: cooking video caption sentence sequence C;
recipe sentence sequence R;
Output: aligned pair list L;
01 generateAlignedPairList(C, R)
02 let t be a threshold value
03 let S be a m x n matrix
04 let c be a m x n matrix // computing the length of an LCS
05 let b be a m x n matrix // simplifying construction of an optimal solution
06 m ← length[C]
07 n ← length[R]
08 call computeCRSimilarity(C, R, S) // caption-recipe similarity
09 t ← threshold // using (1)
11 for each c ∈ C
12   for each r ∈ R
13     if S[c, r] >= t then c[c, r] ← 1
14     else c[c, r] ← 0
15 call LCS-LENGTH(C, R, c, b)
16 call PRINT-AlignedPairList(b, C, R, m, n)
01 computeCRSimilarity(C, R, S)
02 set all elements to 0 in matrix S
03 for each c ∈ C
04   for each r ∈ R
05     S[c, r] ← Sim(c, r) // cosine similarity
01 PRINT-AlignedPairList(b, C, R, i, j)
02 if i = 0 or j = 0
03   then return
04 if b[i, j] = "↖"
05   then PRINT-LCS(b, C, R, i-1, j-1)
06     print "[" + Ci + "," + I + "," + Ri + "," + j + "]"
07 elseif b[i, j] = "↑"
08   then PRINT-LCS(b, C, R, i-1, j)
09 else PRINT-LCS(b, C, R, i, j-1)
```

Figure 4: A caption-recipe alignment algorithm

The algorithm uses sentences of closed captions and recipes as input, and produces aligned pair list. A similarity matrix is constructed by using the sentences. Each element of the similarity matrix is filled up with a similarity value. To measure the similarity value between two sentences, cosine similarity measure for vector space model is utilized. Two sentences are tokenized into words and the words are transformed into vectors separately. Cosine similarity value is measured and the matrix is filled with the values. In similarity calculation using vector space model, word's spelling has a large influence in a similarity value. Measuring semantic relations between two words is also difficult. As a matter of fact, cooking video's caption is based on recipe and terms used to explain cooking equal to terms in recipes. However, specific terms showing numerical value have a bad influence on similarity calculation. For example, an expression "1/2 cup" in a recipe is presented as "half a cup" in a caption. To overcome this weakness, a mapping set is constructed and alignment manager utilizes the mapping set when a similarity value is measured.

While each element of the word-based similarity matrix has discrete value as zero or one, each element of the sentence-based matrix has value between zero and one.

Consequently, the threshold value of cosine similarity should be determined to decide whether a similarity value in an element is valid or not. The threshold value is determined by using (1). Here, avg(sim) means the average value of all cosine similarity values and σ is the standard deviation. α is a parameter and 0.75 is assigned through cross validation.

$$\min = \text{avg}(\text{sim}) + \alpha * \sigma(\text{sim}) \quad (1)$$

Optimal paths in the similarity matrix are detected by applying DP algorithm based on the threshold value. Once caption-recipe alignment process is finished, time-stamp information is added to corresponding step of a recipe and each sentence in the recipe.

Fig. 5 presents an example of caption-recipe alignment results.

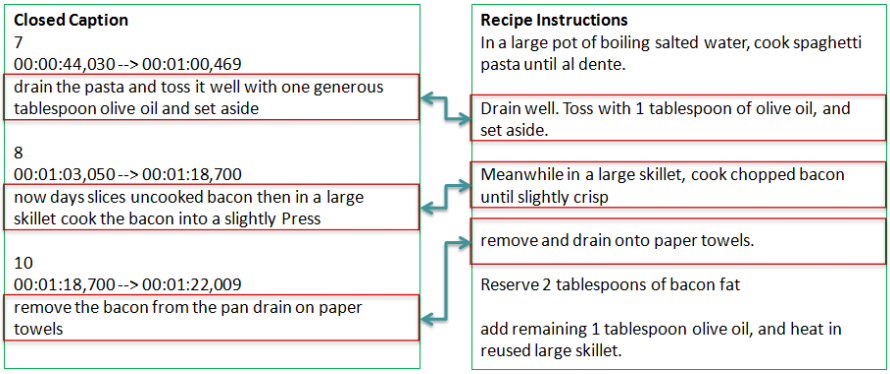

Figure 5: An example of caption-recipe alignment

3.3 Information Extraction using Lexico-Syntactic Patterns

An interaction with an object appeared in a video is an important feature occurred while a user watches the video. Interactive videos should provide relevant information of an object that a user chooses when the user interacts with an object of an interactive video because an interactive video focuses on how to cook a given food using a recipe without explaining all relevant information of objects in the video. Once time information for each instruction of a recipe is identified through caption-recipe alignment algorithm, it is required to extract entity information which is offered to a user when an interaction is occurred. In a case of cooking videos, information such as ingredients, ingredient portion and cooking tools are the information to be offered. Ingredients and cooking tools can be extracted Named Entity Recognition (NER) because those of each instruction consist mainly of noun or noun phrase. Traditional rule-based NER techniques commonly rely on a small set of patterns to identify relevant entities from normal text and they often require gazetteer list for NER. However, they have some limitations from homonym and ambiguity perspectives. LSP is applied for information extraction to overcome the limitations [Jacobs, 91]. A LSP is based on text tokens and syntactic structure of text for a string matching pattern. A task defining pattern using LSP is easier than rule-based NER and alleviates time-consuming for pattern definition. Also, information extraction task

typically achieve a very high precision if LSP is applied to the task [Maynard, 09] [Panchenko, 12].

Table 2 presents some of LSPs used for information extraction. Patterns for chef, ingredient, ingredient portion and cooking tools are defined to extract information.

LSP Category	LSP	Examples
Chef	Recipe by {NNP}* Recipe by Chef {NNP}*	Recipe by Peter Recipe by Chef John
Ingredient & Ingredient Portion	CD pound {VBD}{NN}* CD tablespoon {NN}*	1 pound uncooked spaghetti 1 tablespoon olive oil
Cooking Tool	IN {DT}{JJ}{NN}* IN {DT}{NN}*	In a large pot To pan

Table 2: Examples of lexico-syntactic patterns

3.4 Semantic Entity Interconnection

Interactive videos should provide detailed information about an object such as an ingredient and a cooking tool when a user interacts with an interactive cooking video. Novice cooks have no information on ingredients and cooking tools of a food recipe while experienced and skilled cooks already know the information on them. However, food recipes just describe how to cook the food with given instructions, not provide specific information about entities contained at each recipe instruction. To provide the information within interactive cooking videos, entities should be connected external sources which have detailed information on them. The proposed system provides the information to users by interconnecting entities which are extracted based on LSPs, to Uniform Resource Identifiers (URIs) of DBPedia. DBPedia provides Wikipedia data in linked data format. Wikipedia consists of wiki pages made by crowd-sourcing. A method to interconnect an entity with a DBPedia resource is as follows. Each entity extracted by LSPs consists of a noun or compound nouns. DBPedia can be queried through SPARQL Endpoint [SPARQL Endpoint, 15] with a SPARQL query made by using certain entity name to identify a DBPedia URI for the entity. However, in many cases, SPARQL Endpoint does not return any results because query is not matched to any data of DBPedia. SPARQL query is sensitive at the morphology of the word while entities extracted by LSPs composed of lower cases. To solve this limitation, the proposed approach uses the fact that all entities of DBPedia are created based on each page of Wikipedia and both DBPedia and Wikipedia use same local identifier for two entities. Using local identifier of Wikipedia URL for an entity, existence of the entity on DBPedia is confirmed. A URL of an entity at Wikipedia is preemptively checked instead of running SPARQL query. A user can access wiki page with Wikipedia URL and same problem can be happened like SPARQL query. Wikipedia provides redirection functionality. The redirection function handles capital and small letter, synonym cases and shows a redirected page with a notice that the page is redirected from given URL. Fig. 6 shows an example of this case.

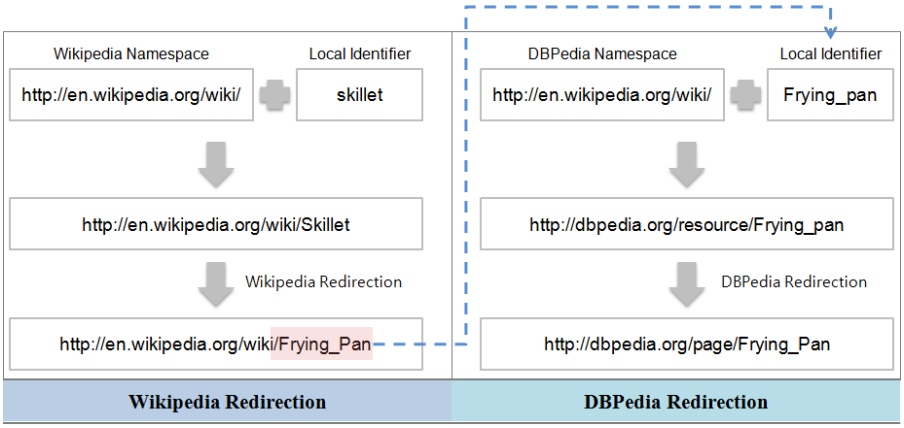

Figure 6: An example of semantic interconnection

When an access of 'skillet' entity using Wikipedia URL is tried, the redirection function changes 'skillet' to 'Skillet' by handling a lower case and shows redirected page for 'Frying_pan' with a redirection notice. Then, an entity name 'Frying_pan' is used to identify DBPedia URI as shown in Fig. 6. Finally, an entity as an object in cooking video such as ingredient and cooking tool is interconnected with the URL and URI. Later, by using both URL and URI, the proposed interactive cooking video system can provide detailed information on objects a user interacts with.

3.5 Cooking Video Annotation and Ontology Population

Cooking video annotation process is performed by generating ontology instances based on CVA ontology and the results of semantic video annotation phase. In order to create ontology instances as annotation data, a set of rule is defined to map html data of recipes and pre-processed data to CVA ontology schema. The set also includes rules to populate the extracted information and semantic interconnection to the schema.

An example of mapping rules is presented as shown in Fig. 7. Left of Fig. 7 shows a rule to define instance name of given recipe. The instance name is determined by using a URL "http://allrecipes.com/recipe/spaghetti-carbonara-ii/". The last part 'spaghetti-carbonara-ii' is used to create the instance name. Right of Fig. 6 presents 'uses' relation between Recipe and Tool classes.

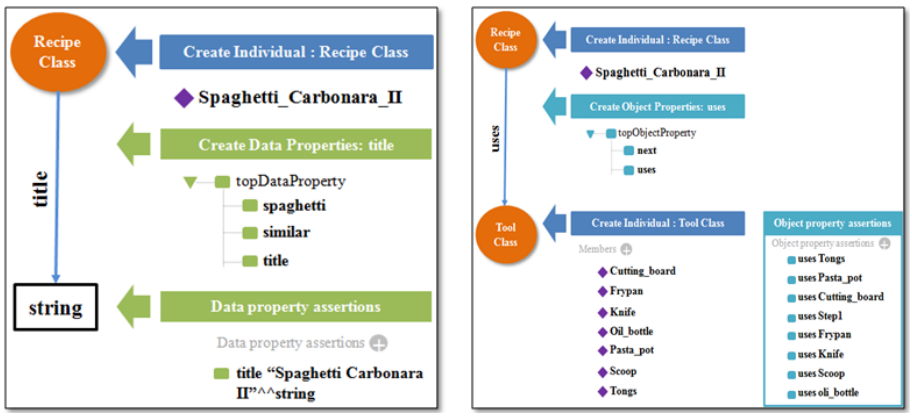

Figure 7: An example of mapping rules

4 Experiments and Discussion

4.1 Experimental Set-up

The proposed approaches for interactive cooking video system requires resources such as cooking videos, food recipes, text captions. To extract these resources, a web crawler is developed by using Java language. Through the web crawler, cooking videos and recipes are extracted from allrecipes.com. Closed captions for cooking videos are generated in SubRip Text (SRT) [SRT, 15] format which is one of caption types, by using YouTube's auto caption feature. To save and manage ontology instances for semantic cooking video annotation using alignment information and recipe data, CVA ontology is applied as a metadata structure. Jena framework [JENA, 15] is used to handle the metadata as ontology instances. Wikipedia and DBPedia as Linked Data [Linked Data, 15] resources are used for semantic entity interconnection. For evaluation of the proposed system, 100 cooking videos are extracted from allrecipes.com. By using 80 cooking videos and relevant recipes among them, the threshold value in the propose alignment algorithm is measured and LSPs for semantic interconnection are defined by analyzing them. 20 cooking videos are used for experiments to evaluate the proposed approaches in the proposed system. Table 3 presents a summary of data set.

"SRT elements" refers to the number of caption elements in cooking videos. "After Refinement" means the number of caption elements after the refinement process. Certain sentence is separated into some SRT elements because extracted closed captions are generated based on time interval between time-stamped tokens.

The refinement process rebuilds the captions with a sentence as a unit. "Word per Sentence" indicates the number of words in a sentence. "Recipe Sentence" and "Words per Recipe Sentence" refer to the number of sentences in a recipe and words in a sentence of the recipe, respectively. Due to speaker's welcome words and

supplementary explanations, there are fewer sentences and words than captions in a recipe.

Factor	SRT elements	After refinement	Words per sentence	Recipe sentence	Words per recipe sentence
average	37	26	14	21	12

Table 3: A summary of experimental data set

According to a research [Liao, 13], the YouTube's Automatic Speech Recognition (ASR) system has a word error rate of about 52% on average over English language videos. They have reduced the average error rate of 44%. Extracted closed captions contain the average error rate of 27%.

4.2 Experimental Results

Interactive cooking video system is proposed to automatically generate interactive cooking videos and to effectively support user interactions occurred during watching cooking videos. To generate interactive cooking videos, caption-recipe alignment algorithm, LSP-based information extraction and semantic entity interconnection method are applied. Experimental evaluation is performed to verify each approach for the proposed system.

Caption-recipe alignment algorithm is evaluated with comparison algorithms associating videos with related text data. The evaluation is performed on 20 cooking videos and food recipes. Table 4 presents alignment accuracy according to each algorithm. Recall, precision and F-measure measurements are used to evaluate each alignment approach. Recall values are measured by calculating the ratio of intersection of actually matched caption-recipe pairs and pairs matched from the proposed algorithm to the matched pairs by the proposed algorithm. Precision values are measured by using the ratio of the intersection to the matched caption-recipe pairs by the proposed system. Lastly, F-measure is calculated as the harmonic mean of recall and precision values.

Approach	Proposed Method	Naïve DP Method	Cooking Navi	Video Cooking
Recall	0.857	0.779	0.757	0.712
Precision	0.89	0.809	0.787	0.793
F-Measure	0.873	0.793	0.772	0.750

Table 4: Experimental results of caption-recipe alignment

Naïve DP method associates recipe instructions to sentences of the closed caption based on words [Turetsky, 04]. Remain process of Naïve DP is performed with a DP method to find out LCS. A method for Cooking Navi associates each instruction of recipe to cooking videos by using total sum of the score derived by three steps [Hamada, 04] [Hamada, 05]. The authors used ordinal restriction by using each

sentence of food recipes. To do this, text blocks of recipe sentences are extracted and ordinal structure of a recipe is analyzed. Text blocks consists of pairs of ingredients and cooking actions. Cooking Navi then extracts video scenes and associate each scene to background such as board, range, table and others. Finally total score is calculated by their proposed equations. The number of the words that are appeared in the text block and the audio data of each video scene, is also used. Video CooKing [Doman, 11] synthesizes cooking recipes to each corresponding video clips. Pairs of ingredients and cooking operations which are commonly appeared in both cooking recipe text and closed captions are extracted. Videos are divided into each clip and cooking operation for each clip is identified. Each pair is tagged to each video clip. The performance of the proposed method outperformed the other systems in terms of all measurements. The analysis of experimental results is described at discussion section.

Table 5 shows accuracy of information extraction based on LSPs. Recall values at here are calculated by the ratio of intersection of actual items and items extracted from the proposed approach to the extracted items by the proposed approach for information extraction. Precision values are measured by calculating the ratio of the intersection to the items extracted from the proposed approach. From each cell of precision, the number of the intersection and extracted items are denoted. For recall, the numbers of the intersection and actual items are presented.

Measurement	**Chef**	**Ingredient**	**Ingredient Portion**	**Cooking Tool**
Precision	20 / 20 (100%)	208 / 227 (91.6%)	219 / 227 (96.5%)	93 / 93 (100%)
Recall	20 / 20 (100%)	208 / 244 (85.2%)	227 / 253 (89.7%)	93 / 105 (88.6%)

Table 5: Accuracy of Information Extraction based on LSPs

LSP-based information extraction produces high precision. Accuracy of ingredient part is the lowest among 4 categories of Table 4. In a case of ‘dry white wine’, ‘dry’ is a part of a wine name while the proposed information extraction method handles ‘dry’ as an adjective. Due to this, accuracy for ingredient is not enough. Information extraction for ingredient portion is worked well by discovering numerical part on ingredient part of a food recipe. Although cooking tools is extracted perfectly, some instance of extraction is duplicated. For example, ‘skillet’ and ‘pan’ is used at a recipe of carbonara. Two cooking tools are different syntactically, but they are same cooking tool in a cooking video. A method identifying whether each extracted cooking tool is same or not in a recipe is required.

Table 6 shows the accuracy of semantic interconnection. Recall and Precision values for semantic interconnection are measured similar to results of information extraction with LSPs. Actual links, generated links by the propose approach, and intersection is used to calculate recall and precision for experiments of semantic interconnection.

Types	Ingredient	Cooking Tool
Precision	73 / 73 (100%)	11 / 11 (100%)
Recall	73 / 76 (96%)	11 / 13 (84.6%)

Table 6: Accuracy of Semantic Interconnection

Because of redirection function of Wikipedia, high accuracy is acquired for semantic interconnection between objects of videos and entities of semantic web.

Through the proposed approaches, interactive cooking videos are generated. Most of the recipe instructions are associated with corresponding part of cooking videos by using caption-recipe alignment algorithm. Main entities which are the medium for user interactions are extracted through LSPs. Objects of cooking videos are interconnected with semantic web entities. Semantic annotation of cooking video is performed based on CVA ontology and the information such as alignment and LSP-based extraction.

Fig. 8 presents a user interface which plays an interactive cooking video generated from the proposed system.

Figure 8: An example of semantic cooking video annotation

User can navigate cooking instruction by clicking next and previous buttons. At bottom-right corner, a speech bubble is located to provide the toggle function for showing the current recipe step. The user interface also allow user to easily find a specific point which is a certain instruction of a recipe a user want to watch by clicking the buttons. This is available because the proposed caption-recipe alignment process is performed by sentences and the timeline of text caption is interconnected with each sentence of recipes. Users also can interact with the proposed interactive cooking video system to get extra information about objects in cooking videos. If a user clicks ingredient or cooking tool icon at the top-left corner, the user can obtain

related information in current context without any efforts finding the related information from a search engine with keyword-based search.

4.3 Discussion

The proposed system is superior to the other systems in caption-recipe alignment perspective. Through the analysis of experimental results, we reveal the reason that the other methods produced the lower performance than the proposed method. Naïve DP method constructs a similarity matrix based on words and finds an optimal path. An optimal path should form diagonal line near to 45 degree angle to obtain better alignment performance. However, difference between declarative sentences of the captions and imperative sentences of recipes produces an optimal path which has serpentine course. In addition, alignment performance is reduced because incomplete captions have a negative impact on building an optimal path. A method for Cooking Navi uses combined information derived from ordinal structure of a recipe, shot classification and co-occurrences of words in a caption and a recipe. However, ordinal structures not well build in some recipes give bad effects. Shot classification and verbs in instruction is associated to four types: board, table, range and others. This dimension reduction causes information loss and alignment performance is deteriorated. The method of Video CooKing produced the lowest alignment performance despite the synonym problem of cooking operation is solved by applying word similarity with WordNet [Pedersen, 04] in experiments at here. In cases of Cooking Navi and Video CooKing, they used pre-established closed captions in their system. However, at here, closed captions are not perfect and some misrecognized words have a bad influence on the alignment process.

The proposed caption-recipe alignment algorithm can achieve higher accuracy than benchmark algorithms through elimination of winding path by applying sentence-based approach and information loss caused by dimension reduction.
Cooking Navi and Video CooKing have tried to provide interactive cooking video services, but their systems just perform alignment process. They still do not provide another user interaction with objects appeared in cooking videos as mentioned in Interactive Video section. However, the proposed system shows the better performance from the alignment perspective and allows users to interact with interactive cooking videos generated.

5 Conclusions and Future Works

Videos become very popular resources in the Web and handling a huge volume of video data become a very important task in information retrieval field. People not only watch videos but they also want to interact with videos to find a specific part or get detailed information on objects in videos. To support user interactions, current videos should be transformed into the forms of interactive videos.

In this paper, an interactive cooking video system is proposed to generate interactive cooking videos and to support user interactions. To do this, the system automatically synchronizes each recipe instruction to each part of cooking videos through the proposed caption-recipe alignment algorithm. The system extracts ingredient and cooking tool information by using LSPs from each instruction to

support a user interaction with objects of videos. Each entity as the extracted information is interconnected with Wikipedia entity as things of semantic web. Finally, the proposed system automatically generates interactive cooking videos by semantically annotating cooking videos with modeled CVA ontology.

To evaluate the proposed system, the accuracy of each approach used in the system is measured. For an accuracy of caption-recipe alignment algorithm, a comparative experiment is performed. Experimental results show that the proposed method is superior to compared algorithms in alignment accuracy perspective. LSP-based information extraction method achieved high accuracy over 95% for given food recipes of cooking videos. Semantic entity interconnection method also produced high accuracy for interconnecting ingredients and cooking tools. The user interface allows a user to easily find specific scenes within a cooking video and to acquire detailed information on entities of a cooking video.

From experimental results, the proposed method shows satisfactory results on the alignment, information extraction and semantic entity interconnection. However, the performance on the proposed caption-recipe alignment algorithm still can be improved by applying weighting techniques [Guo, 12] [Liu, 13] [Saric, 12]. In addition, based on the generated interactive cooking videos, an interactive video service like [Hamada, 04] [Schafer, 13] can be provided at cooking video web sites. From a usage perspective, the proposed caption-recipe alignment algorithm can be used to improve closed captions of cooking videos.

This research is not limited to only cooking video domain. In the future work, the proposed approach will be applied to how-to videos which have procedural knowledge within videos similar to cooking videos. Metadata obtained from how-to videos by using the proposed methods can be used to generate interactive how-to videos which support user interaction be occurred during watching videos and construct an interactive how-to video system. With generated interactive how-to videos, the system can provide automatic illustration of recipes to corresponding video scenes. In addition, it is possible to offer how-to video search and search within a video.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No. 2015-R1A2A2A03006190)

References

[Allrecipes, 15] Allrecipes, accessed Sep 2015: http://allrecipes.com/

[Ballan, 11] Ballan, L., Bertini, M., Bimbo, AD., Seidenari, L., Serra, G.: Event detection and recognition for semantic annotation of video, Journal of Multimedia Tools and Applications, 51(1), 279-302, 2011.

[BBCFood, 15] BBC Food, accessed Sep 2015: http://www.bbc.co.uk/food/

[Bellman, 09] Bellman, S., Schweda, A., Varan, D.: A Comparison of Three Interactive Television AD Formats, Journal of Interactive Advertising, 10(1), 14-34, 2009.

[Cour, 08] Cour, T., Jordan, C., Miltsakaki, E., Taskar, B.: Movie/script: Alignment and Parsing of Video and Text Transcription, In Proc. of the 10th European Conf. on Computer Vision: Part IV, October 2008, 158-171.

[Doman, 11] Doman, K., Kuai, CY., Takahashi, T., Ide, I., Murase, H.: Video CooKing: Towards the Synthesis of Multimedia Cooking Recipes, The 17th International Multimedia Modeling Conf, January 2011, 135-145.

[Foodista, 15] Foodista, accessed Sep 2015: http://www.foodista.com/

[Guo, 12] Guo, W., Diab, M.: A Simple Unsupervised Latent Semantics based Approach for Sentence Similarity, In Proc. of First Joint Conf. on Lexical and Computational Semantics, June 2012, 586-590.

[Hamada, 04] Hamada, R., Miura, K., Ide, I., Satoh, S., Sakai, S., Tanaka, H.: Multimedia Integration for Cooking Video Indexing, In Proc. of Pacific-Rim Conf. on Multimedia, December 2004, 657-664.

[Hamada, 05] Hamada, R., Okabe, J., Ide, I., Satoh, S., Sakai, S., Tanaka, H.: Cooking Navi: Assistant for daily cooking in kitchen, In Proc. 13th annual ACM International Conf. on Multimedia, November 2005, 371-374.

[Homer, 14] Homer, BD., Plass, JL.: Level of Interactivity and executive functions as predictors of learning in computer-based chemistry simulations, Journal of Computers in Human Behavior, 26, 365-375, 2014.

[Jacobs, 91] Jacobs, PS., Krupka, GR., Rau, LF.: Lexico-Semantic Pattern Matching as a Companion to Parsing in Text Understanding, In Proc. of the Workshop on Speech and Natural Language, collocated with the 6th Human Language Technology Conf., 1991, 337–341.

[JENA, 15] Apache Jena Framework, accessed Sep 2015: https://jena.apache.org/

[Lambert, 13] Lambert, A., Guegan, M., Zhou, K.: Scene Reordering in Movie Script Alignment, In Proc. of 11th International Workshop on Content-based Multimedia Indexing, June 2013, 213-218.

[Liao, 13] Liao, H., McDermott, E., Senior, A.: Large Scale Deep Neural Network Acoustic Modeling with Semi-Supervised Training Data for YouTube Video Transcription, In Proc. of IEEE Automatic Speech Recognition and Understanding Workshop, December 2013, 368-373.

[LinkedData, 15] Linked Data – Connected Distributed Data across the Web, accessed Sep 2015: http://linkeddata.org/

[LinkedRecipes, 15] Linked Recipe Data, accessed Sep 2015: https://code.google.com/p/linkedrecipes/

[Liu, 13] Liu, Y., Liang, Y.: A Sentence Semantic Similarity Calculating Method based on Segmented Semantic Comparison, Journal of Theoretical and Applied Information Technology, 48(1), 231-235, 2013.

[Maynard, 09] Maynard, D., Funk, A., Peters, W.: Using lexico-syntactic ontology design patterns for ontology creation and population, In Proc. of the Workshop on Ontology Patterns, collocated with the 8th International Semantic Web Conf., October 2009, 39-52.

[Panchenko, 12] Panchenko, A., Morozova, O., Naets, H.: A Semantic Similarity Measure Based on Lexico-Syntactic Patterns, In Proc. of the 11th Conf. on Natural Language Processing (KONVENS), September 2012, 174-178.

[Pedersen, 04] Pedersen, T., Patwardhan, S., Michelizzi, J.: WordNet::Similarity – Measuring the Relatedness of Concepts, In Proc. of Human Lanugage Technology/North American chapter of the Association for Computational Linguistics annual meeting, May 2004, 38-41.

[PopcornMaker, 15] Popcorn Maker, accessed September 2015: https://popcorn.webmaker.org/

[Ronfard, 03] Ronfard, R., Tran Thuoug, T.: A framework for aligning and indexing movies with their script, In Proc. of IEEE International Conf. on Multimedia and Expo (ICME), July 2003, 21-24.

[Saric, 12] Saric, F., Glavas, G., Karan, M., Snajder, J., Basic, BD.: TakeLab: Systems for Measuring Semantic Text Similarity, In Proc. of First Joint Conf. on Lexical and Computational Semantics, June 1021, 441-448.

[Schafer, 13] Schafer, U., Arnold, F., Ostermann, S., Reifers, S.: Ingredients and Recipe for a Robust Mobile Speech-Enabled Cooking Assistant for German, In Proc. of 36th annual German Conf. on AI, September 2013, 212-223.

[SPARQLEndpoint, 15] SPARQL Endpoint, accessed Sep 2015: http://www.w3.org/wiki/SparqlEndpoints

[SRT, 15] SubRip Text File Format, accessed Sep 2015: https://en.wikipedia.org/wiki/SubRip

[Turetsky, 04] Turetsky, R., Dimitrova, N.: Screenplay Alignment for Closed-System Speaker Identification and Analysis of Feature Films, In Proc. of IEEE International Conf. on Multimedia and Expo, June 2004, 1659-1662.

[Vimeo, 15] Vimeo, accessed Sep 2015: https://vimeo.com/

[Wirewax, 15] Wirewax, accessed Sep 2015: http://www.wirewax.com/

[Wu, 12] Wu, J., Worring, M.: Efficient Genre-Specific Semantic Video Indexing, IEEE Transactions on Multimedia, 14(2), 291-302, 2012.

[YouTube, 15] YouTube, accessed Sep 2015: https://www.youtube.com/