I J C T A, 9(15), 2016, pp. 6993-7005

Automatic Source of Domain Module Gathering Information From Electronic Document

Amandeep Singh* and K. Deeba**

ABSTRACT

Automatic keyword is theprocess in which the keyword is automatically extracted from a text document. Here we try to presenting a content based system for automatically keyword extraction and recommendation. Building the Domain Module is a hard task which entails not only selecting the domain topics to be learned and also defining the complex problems among the topics that determine how to plan the learning sessions. Textbook authors deal with similar problems while writing their documents, which are structured to facilitate comprehension an Electronic textbooks might be used as the source to build the Domain Module, reproducing how average teachers behave while preparing their subjects.Stemming and frequency Algorithm gathering - At this phase, the domain topics to be performed in complex problems among them are identified and represented in the LDO. The LDO will allow either the Stemming to plan the learning session or the students to guide themselves during the learning process.

Keywords: Data modelling, Domain Modules, Electronic Document, DOM Sortze framework

1. INTRODUCTION

1.1. General Introduction

The revolution of information and communication technologies has affected education providing means to enhance both the teaching and learning process.

The technology supported learning systems (TSLS), such as intelligent tutoring system(ITS), adaptive hypermedia system (AHS) and especially learning management system (LMS) such as Moodle1 or Black board are being widely used in many academic institutions and becoming essential for process.

Further positive relationship between the usage in Web based learning technology and student engagement and desirable learning outcomes has been designed.

To be effective TSLS require an appropriate Module where the pedagogical representation of the domain to be learned.

The Module is considered the core of any TSLS as it represents the knowledge about a subject matter to be communicated to the process. An incomplete Module may result in a system that is only able to provide information of the instruction required for the domain process.

Building the Module is a hard task required by not only selecting the domain topic to be learned but also defining the relationship among the topics that determine how to plan the learning session.

Text book author deal with similar problem while writing their documentation which are structured to facilitate comprehension and learning.

* Post Graduate Student, *Email: aman27892@gmail.com*

** Associate Professor Department of computer Science and Engineering SRM University.

Electronic textbook might be used as the source to build the Module reproducing how average teachers behave while preparing their subjects they choose a set of reference books that provide the main didactic resource (DR) for the subject and rely on them for scheduling their lectures.

Artificial intelligence techniques provide the means for the semiautomatic construction of the Domain Modules from electronic textbook which may significantly contribute to reduce the development cost of the Domain Module process.

This paper presents Sorted a framework for the semiautomatic type generation of the Module from electronic textbook Sortze aims to be domain dependent so that no domain-specific knowledge is used for excepting the processed electronic textbook and the knowledge gathered from it and describes the Module generation process which entails three main tasks such as preparing the document for knowledge extraction process.

Building the ontology that describes the domain to be learned and the generation of the learning object (LO) presents a real case environment in which the work here described has been applied to be evaluated.

Finally the conclusions and future work are mentioned in theprocess .

2. DOMAIN MODULE GENERATION

The semiautomatic generation of the Domain where work of the Module encodes knowledge at two different levels such as the learning domain ontology (LDO) and the set of LO.

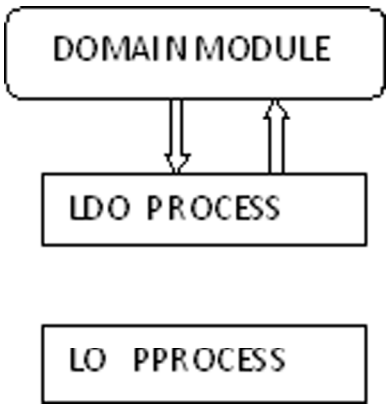

Figure 1: Representation of module

The following steps are carried out to develop and design the Module structure as:

1. Textbook preprocess method: First the document must be prepared for the subsequent knowledge based acquisition process. This process is described here and the outcomes are then used to gather the two levels of knowledge encoded in the Module.

2. LDO gathering: At this phase the domain topics to be mastered and as well as the pedagogical relationships provided by it among them it is identified and represented in the LDO process.

The LDO will allow it s part by gathering TSLS to plan the learning session orthe students to guide themselves during the learning process and it s developed further.

The acquisition of the LDO is described in this process.

3. LO gather processing: At this stage the LO-definitions, examples,development and so on to be used during the learning process are identified and generated during process of domain purpose.

In this semiautomatic approach, the outcome of gathering the LDO and the LO can be supervised by teachers and instructional designers.

Both individually and collaboration using DOM-Sortze a concept-map-based tool for the supervision of the Domain Module authoring process.

Teachers could, this way for adapting the resulting Domain Module to their requirements or teaching preferences.

Next each of the step is described in detail of domain and the work here described has been applied on electronic documents written in the Basque language.

As part of implementation but for the sake of readability where the examples will be shown in both Basqueand English although some information might be lost in translation process required and it s been established here.

3. TEXTBOOK PREPROCESS METHOD

In this phase all the system prepares the electronic document and gathers a standardized representation of it to later run the knowledge and formation of acquisition processes.

As an electronic documents are available in many different formats such as pdf, rtf, doc thereby the preprocess is carried out first to prepare the document.

The content of electronic documents isorganized using a hierarchical structure by which documents contain chapters, which in turn are divided into sections and so on.

A tree-likeinternal representation of the document is built and so that the rest of the task can be performed with no dependence on the format the original document is stored in.

In addition with the outline of the document, which might be located either at the beginning or the end of the document canalso be numbered or figured out by the process.

The process is indented in different ways showing its structure. Thus homogenized internal representation of the outline is also gathered in thepreprocess method.

The obtained internal representations for the outline and the document body are then systemically analyzed to enhance them with the part-of-speech information that will be used in the following steps determined Systematic analysis is especially for agglutinativelanguages such as Basque language where most words are formed by joining morphemes method indicated together.

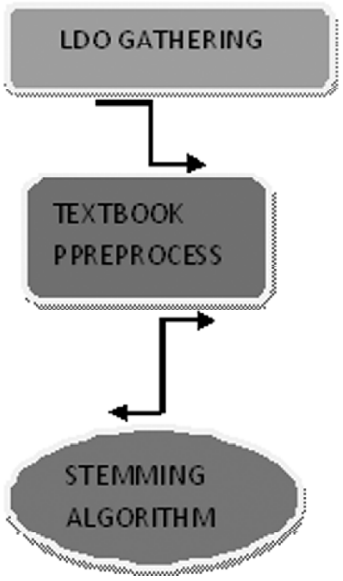

Figure 2: Textbook preprocess method

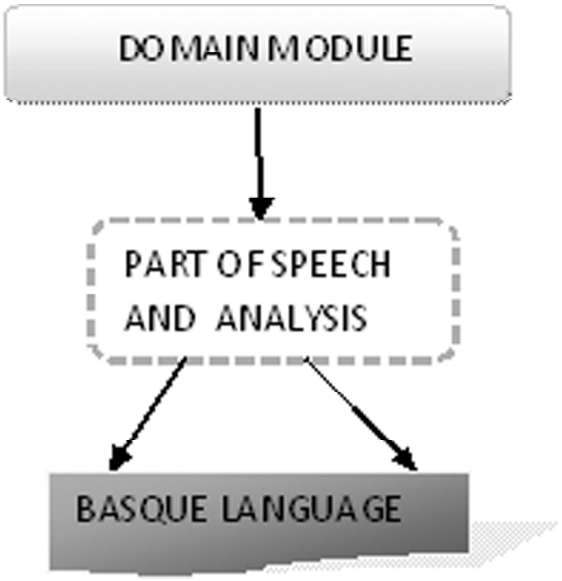

Figure 3: Spefication in Basque language

In the Basque language for example words are formed by adding the affixed to the dictionary entries.

More specifically, the affixed corresponding to the determined number and declension case are taken in this order, independently of each other order.

As prepositional functions are realized by case suffixed inside word forms Basque presents a relatively high power to generate inflected word forms which makes morphosystematic analysis very important to be able to extract information from text fragments.

4. GATHER THE LDO PROCESS

The Ontology learning and gathering domain ontologies from different resources in an automatic system or semiautomatic way has been addressed in many works.

Most of these projects aim at building or extending a domain ontology or populating the analysis of lexical ontologies such as Wordnet development.

Ontology learning usually combines machine learning and NLP techniques to build domain ontologies development or to enhance and populate some base ontologies system.

Different kinds of resources such as document warehouse process, machine readable dictionaries or lexical ontologies such as systematically are broadly used as sources of information for ontology learning.

In the approach presented, the LDO contains the main domain topics to be performed and the pedagogical and systematic relationships along them that are carried by it.

The LDO acquisition entails two main NLP and heuristic relationship based steps are outline analysis which can results in an initial version of the LDO gathering and the document body analysis which enhances the ontology with new topics and relationships are implemented by it.

During the LDO gathering process, an internal representation is used and the representation besides the learning topics and the relationships

information about the gathering process itself used heuristics relationship confidence on the heuristics design and so on is also included.

Once the LDO has been gathered and reviewed by teachers or instructional designers the ontology is represented in OWL where it can be predicted.

4.1. Outline Analysis procedure Document outlines

Are the main sources of information for acquiring the LDO in a semiautomatic way and as they are usually well structured and contain the main topics of the domain.

Besides they are considerably summarized and therefore meaningful information can be extracted with a low-cost process that are indicated.

The reason is that authors of textbooks have previously analyzed the domain and decided how to organize the content according to pedagogical and systematic principles.

Provided that the organization of thetextbook is reflected in the outline of NLP techniques and a collection of heuristics relationship are used to the implicit the function.

The outline analysis is composed of two process they are Basic analysis and In this task the main topics of the domain and the relationships among these topics are mined from the homogenized relationship outline internal and external representation.

Besides the structure of the document outline is used as a means to gather pedagogical relationships that occurs.

A subitem of a general topic is used to perform an certain part of it or a particular case of it.Therefore are defined by every outlineitem procedure and all subitems of entity carried.

In addition the order of the outline items reflects the recommended sequence for learning the domain topics proceed

Thus an initial set of sequential relationships is identified from the order of the outline items reflected by it.

4.2. Heuristics relationship analysis indication

The results of the basic analysis is refined based on a consequence of heuristics that bothcategorizes the relationships identified in the previous step and also mine new ones that define mainly prerequisite relationships among them.

The identified relationships are labeled with the inferred kinds so that the heuristic are used and the confidence on the inferred information are recovered by it.

The heuristics entails the condition to be matched the empirically gathers confidence on the heuristic reasoning and the postcondition of the relationships that are recognized.

The heuristic analysis method is carried out in two steps thereby first the heuristics for identifying structural integrity relationships are applied

A set of heuristics must be defined to perform this analysis performance.

Some of the heuristics are language dependent as they rely on systematic structures that may vary depending on the language they are defined for. As mentioned above this work has been applied and studied on specific documents in the Basque language purposed by it.

This study is allowed for the identification of the set of heuristics and their confidence level .

The following procedure was carried out to identify the heuristics related Analogy:

1. A small set of outlines related to computer science development was analyzed to detect some patterns that might help in the classification of relationship carried.
2. These heuristics weretested on a wide set of outlines related to different domains system
3. The relationships identified by the heuristics were contrasted with the real ones that is manually labeled relationships.
4. After analyzing the results, paying special attention to the detected tasks in the heuristics some new heuristics were defined by it.

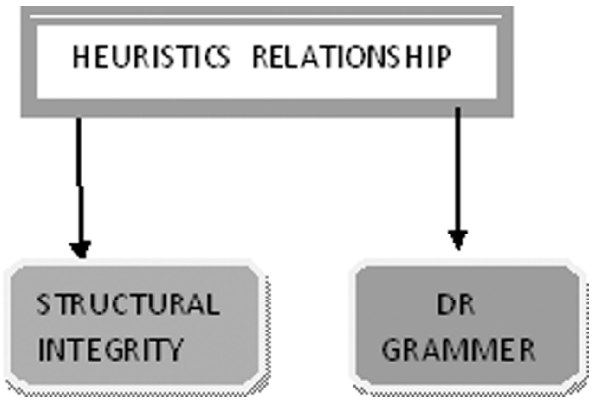

Figure 4: Heuristics analogy

A set of heuristics relationship must be defined to perform this type of analysis.

Some of the heuristics related are being language dependent asthey rely and purpose on syntactically structures that may vary depending on the language they are defined for it. As mentioned above this workhas been applied on documents in the Basque language analysis.

The set of heuristics was identified on a set of outlines purpose of different subjects at the University of the Basque Country.

Their study allowed the identification of the set of heuristics related and their confidence level and the following procedure was carried out to identify the heuristics related implementation.

1. A small set of heuristics outlines related to computer science field was analyzed to detect some patterns that might help in the classification of relationship and implementation of outlines.
2. These heuristics were tested on a wide range of set outlines related to different domains structure
3. The relationships identified by the heuristics implemented were contrasted with the real ones like manually labeled relationships.
4. After analyzing the results and paying special attention to the detected lacks in the heuristics related to some new heuristics is defined.

4.1.1. Heuristics implementation for Structural integrity Relationships

The heuristics related for structural relationships allow identifying the kind of relationship between an item of the outline structure and its subitems.

The heuristic analysis works under the assumption that only one kind ofstructural relationship can exist between an outline structure item related and all its subitems as this fact is observed in almost all the analyzed procedure outlines structure .

The analysis of the outlines also showed that the most common structural relation is the observed relationship. In addition with some homogeneous structures is observed of identified as relationships.

A set of group heuristics observed in heuristics that check if the outline item structure in all its subitems meet a particular condition that allow8 recognizing such heuristics relationships are defined.

Individual heuristics is observed that check whether a particular subitem meets a condition is defined for identifying structural relationships in outline items with heterogeneous and external subitems observed.

The heuristics related that check whether a particular subitem of a general meets a condition is defined for identifying structural and combined with relationships in outline items with heterogeneous subitems of the structured relationship.

The following process is extended for the identification of the structural relationship carried with

1) If a group of heuristic trigger structured with relationships are defined between the outline item and all its subitems
2) Thereby every subtitem an individual heuristic that matches is looked up and for the case where several heuristics relationship could be applied with most confident one is returning the default heuristic value when no other heuristic condition is carried by it

4.1.2. Individual behaviour ofstructural heuristics method

The heuristics check if an individual subitem meets a condition process and it involves thegeneral item.

The empirical analysis shows the different heuristics of this kind can triggered together in the same group of subitems relationships

1. Multiword heuristic structure: The Multiword terms contain information to structure and carried with pattern has been used to gather.
2. Entity name heuristic relationship: Entity names are used to identify examples of a particular entity relationship so thereby the subitems containinformation

4.1.3. Group and systematic structural heuristics

The individual structural heuristics test on outline item and particular subitem match such a kind ofcertain condition in structural relationship where the group structural heuristics check whether the general item or all its subitems match acondition.

Two heuristics of this kind have been identified are:

1. Keyword heuristic relationship: The heuristic rely on a entity of identify relationships among an outline structure of topic and its subtopics.The set of keywords is identified as the set of outlines structure analyzing to define the heuristics method and stored in a configuration file and modified.
2. Common head þ multiword structure heuristic relationship: The heuristic checks whether the subitems of an outline structure item share a homogeneous related issues.

4.1.4. Document structure of Body Analysis

In particular stage of the LDO is enhanced with new topics and relationships gathered from the document body. To achieve this process of goal two processes are generated out. first the new topics are identified as described and next new pedagogical relationships among the topics is identified into process.

5. IDENTIFY NEW RELATIONSHIP AMONG GENERTED TOPICS

The process allows the identification of new pedagogical relationships from the electronic document developed using a pattern- based approach.

Thereby the patterns recognize pedagogical relationships between domain topics structure based on the syntactic appeared in sentences. Therefore the internal outline representation of the document is applied by label any domain topic.

Then nested domain topics such as domain topics constructed on other domain topics is as identified to propose relationships among them.

The grammar contains an entity set of rules describing syntactic structures correspond for pedagogical relationships among them.

The grammar for identifying pedagogical relationships entails rules for recognizing structural relationship and the prerequisite sequential relationship.

The rules that defined after an empirical analysis of a set of textbooks corresponding to structure of an prerequisite are defined in the grammar. When two domain topics are related by the structural expression an IsA combined relationship between these topics can be inferred.

5.1. Gathering LO analysis From electronic Documents

The generation of LO for the domain topics are achieved through identified structureor gathering of DR. The consistent fragments of the document structure with outline related to one or more topics with a particular case.

The identification and extraction of these pieces is carried out in an ontology carried driven by process that also uses NLP techniques.

As the LOgenerating approach presented in work aims to be domain independent where only domain specific knowledge part used is the LDO that has been gathered from the electronic document phase.

The document meant to be used in the learning sessions while a LO refers to a reusable DR enriched with Metadata process and The LO generation process here described by it.

The grammar for gathering the DRs from the electronic document has generated and developed using the Constraint structure of Grammar formalism. The DR grammar is performed on electronic textbooks to observe its performance and reliability.

Some of the initially rules are removed from the final version of the DR grammar Thereby precision of the grammar rules is used to determine the confidence on rules.

The identified DR contain the sentence that triggered the rule for the corresponding DR structure and all the sentences that follow which refer to the same topic where Every DR is labeled with the domain topics referred by it.

The DR is identified by DR grammar are usually quite simple and they are enhanced to make them more accurate. On the performance wise the consecutive DR are combined to which end similarity measures have been defined.

On the other purpose the cohesion of the DR is previously fragmented added to each DR if it contains references and previous DR to sentences. The composite DR is built as an aggregation of DR of lower granularity and keep the information constraints

6. COMPOSITION OF CONSECUTIVE TYPE SIMILAR DR STRUCTURE

The composition of consecutive is reusable to DR and based on the similarity between the DR structure which is determined by aspects The methods that determines the similarities return a value like (0,1) range. Structure is considered in similar corresponding structured values.

Thereby different similarity measuring methods were defined and tested as the ontology based methods were those which provided the most accurate results accessed by it. The ontology topics considering both the semantic relatonships in the ontology format and the topics gathered in the analyzed fragment.

6.1. From DR structure to LO analysis

The possibility to retrieve the desired LO from a large set is a combined key issue to promote the usuage ofLO. The selection of a suitable LO is highly influenced on the metadata process.

While the manual creation of metadata can be considered for annotation of a single LO and therefore the semiautomatic metadata generation can overcome metadata inconsistency issues. The presentation format of the LO may also be affected and not appropriate for flexible content reuse performed as the components cannot be easily accessed easily

The automatic Metadata process the generator and the metadata is enhanced with more information that is extracted during the DR generation for improved Analogy. Most keyword annotation applications use and perform statistical analysis methods and rely on the frequency of the terms in the analyzed text.

The LDO is identified as domain topics in the LO are used to get a more accurate keyword list as the semantics of the relationships is taken into account where a content model for the LO and its component of an reference.

The learning resource type like specified in terms of Analogy which represents a content model for the LO structure and its components.To determine an Learning Resource Type which executes the rules of the DR grammar rules which identify different kinds of precision of DR.

6.2. LO structure vs Storage

The LO previews files have been generated where they are inserted in the LOR to allow their retrieval and use for performance of DR. thereby LO is compelled by all its components are also appropriately labeled and stored as action.

7. EXPERIMENT ANALYSIS AND RESOURCES

The main goal of mining experiment where to evaluate DOM Sortz module where it tends the teachers tobuild the Module by processing the knowledge in the LDO and LO gathered from the textbook sources. the experiment process is on LOwhere the electronic textbook in which the images is processed.

To evaluate the process of generation of the Module using DOM Sortze module by combining a reference LDO and DR results.

The instruction values generated relevant domain topics and the pedagogical relationships to define the LDO structure.

The generation of LDO is evaluated and processed based for automatically gathered knowledge is gathered among domain topic and pedagogical relationships. The evaluation of the LO generation is now being considered as the identified LO combined to DR.

7.1. Process of Gathering LDO method

The evaluation of the LDO is carried out by comparing the identified domain and pedagogical relationships to the reference LDO structure.

The LDO elements levels is constraint in domain topics and the pedagogical relationships.The outline isanalyzed to gather the initial LDO for document body is processed to identify new topics structure and relationships.

Figure 5: LDO Storage

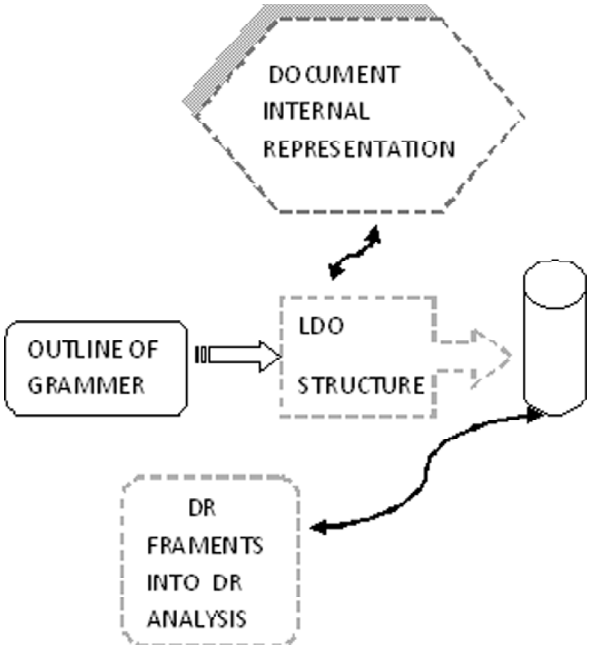

Figure 6: Estimation of DR structure

7.2. Estimation of the Gathered structure LDO process

The estimated of the construction of the LDO is carried out comparing the domain topics and pedagogical relationships to the LDO structure. The Elimination of the LDO is carried out in two steps

1. The outline is analyzed to gather LDO
2. The document body is to identify new topics structure and relationships Analogy.

The process of the LDO identification is achieved by the precision method. The precision method to facilitate pedagogical relationships were identified. The identification of the pedagogical relationships is been measured as source of relationship.

This is mainly due to domain specific knowledge to determine relation between structure of topics. Even though some systematic patterns were defined to knowledge source. The heuristics relationships for the analysis of the outline structure achieved by higher performance used in the analysis of the whole document body.

7.3. Evaluation of Gathered processing LO Analogy

LO acquired is to assess an particular context while one of its components is compared to DR grammar.

The DR grammar is evaluated to determine its accuracy and it s process method the gathered LO is identified set of DR to get the percentage of automatically gathered DR and their correctness or accuracy value combined with it.

8. EVALUATION OF GATHERED PROCESSING FOR DR GRAMMAR

The atomic structure of LO is the finer grained LO built from the identified atomic DRgrammar have been checked and executed for means to assess the pattern based approach for gathering process from the electronic textbookscontained validation of DR fragments.

These patterns make easier problem related statement easier to find and provide better results than the rules for other kinds of DR. Therebyexamples and kinds of DR rely on patterns requirements or the sentence

might correspond to another kind of DR rather than that identifiedprocess so they form a lower reliability consequences.

9. EVALUATION OF THE LO GATHERED PROCESS

The gathered LO is evaluated to determine identified DR with the automatically gathered ones to measure and analyze the gathered LO.

Many of the manually identified DR structure is an composite fragments that contain structure resources.Although the instructional structure identify any principle or example as they require components of the composite LO Analogy.

10. CONCLUSION AND FUTURE ENHANCEMENT

This paper is presented an Module DOMSortze for the semiautomatic generation of the Modulefrom electronic document and sources combined with textbooks.

The system requires techniques for heuristic reasoning and ontologies for theknowledge acquisition processes.

DOMSortze is tested using an electronic textbook and comparing the elements with Domain name. The procedureto evaluate DOMSortze contributes to Domain Module authoring process.

The rules for defining performed better action probably because the sentences employed in the book are shorter and less complex structureanalysis.

The identification of problem related statements in Basque is facilitated by an auxiliary language. Future enhanced of DOM Sortze is planned to enhance grammar for identifying pedagogical relationships acquiring the recall of the relationships.

Athough DOM Sortze is able to process images in the electronic document the image is referenced and thus treatment of images must be improved.

The LDO ontology as the domain might being to get approximate translations of the gathered LO used for searching and retrieving information.

Thereby machine learning methods are planned for new rules that might improve the identification at the DR in the electronic document structure textbooks.

The construction of this model is semi-automated so that the development efforts from developers can be reduced.

The user-profile learning algorithm responsible for expanding and maintaining up-to-date the long-term user’s interests, employs a domain-based inference method in combination with other relevance feedback methods to populate more quickly the user profile and therefore reduce the typical cold-start problem.

The filtering algorithm, which follows a stemming approach making usage of a semantic similarity method based on the hierarchical structure of the ontology to refine the item-user matching score calculation.

11. FUTURE WORK

Future work on DOM Sortze comprises improving the generation of the LDO. It is planned to enhance the grammar for identifying pedagogical relationships to increase the recall of the relationships.

Alternative ways to gather prerequisite relationships which have a very poor recall will be also tested.

REFERENCES

[1] C.F. Chai and C. Hung, "Automatically Annotating Images with Keywords: A Review of Image Annotation Systems," Recent Patents on Computer Science, vol. 1, pp. 55-68, 2008.[20] J.-Y. Pan, H.-J. Yang, and C. Faloutsos, "MMSS: Multi-Modal Story-Oriented Video Summarization," Proc. Fourth IEEE Conf.

[2] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," Machine Learning, vol. 41, no. 2, pp. 177-196, 2001.

[3] F. Monay and D. Gatica-Perez, "Modeling Semantic Aspects for Cross-Media Image Indexing," IEEE Trans. Pattern Analysis and Machine Intelligence.

[4] T.L. Berg, A.C. Berg, J. Edwards, and D. Forsyth.

[5] M. Ozcan, L. Jie, V. Ferrari, and B. Caputo, "A Large-ScaleDatabase of Images and Captions for Automatic Face Naming," Proc. British Machine Vision Conf., pp. 1-11, 2011.

[6] J. Luo, B. Caputo, and V. Ferrari, "Who's Doing What: JointModeling of Names and Verbs for Simultaneous Face and Pose Annotation," Advances in Neural Information Processing Systems,vol. 22, pp. 1168-1176, 2009.

[7] J. Wang, K. Markert, and M. Everingham, "Learning Models for Object Recognition from Natural Language Descriptions," Proc. British Machine Vision Conf., 2009.

[8] S. Ju Hwang and K. Grauman, "Learning the Relative Importance of Objects from Tagged Images for Retrieval and Cross-ModalSearch," Int'l J. Computer Vision, pp. 1-20, 2011.

[9] V.O. Mittal, J.D. Moore, G. Carenini, and S. Roth, "Describing Complex Charts in Natural Language: A Caption Generation System," Computational Linguistics, vol. 24, pp. 431-468, 1998.

[10] M. Corio and G. Lapalme, "Generation of Texts for Information Graphics," Proc. Seventh European Workshop Natural Language Generation, pp. 49-58, 1999.

[11] L. Zhou and E. Hovy, "Headline Summarization at ISI," Proc. HLT-NAACL Text Summarization Workshop and Document Understanding Conf., pp. 174-178, 2003.

[12] R. Soricut and D. Marcu, "Stochastic Language Generation Using WIDL-Expressions and Its Application in Machine Translation and Summarization," Proc. 21st Int'l Conf. Computational Linguistics and the 44th Ann. Meeting Assoc. for Computational Linguistics, pp. 1105-1112, 2006.

[13] S. Wan, R. Dale, M. Dras, and C. Paris, "Statistically Generated Summary Sentences: A Preliminary Evaluation of Verisimilitude Using Precision of Dependency Relations," Proc. Workshop Using Corpora for Natural Language Generation, 2005.

[14] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees," Proc. Int'l Conf. New Methods in Language Processing, 1994.

[15] Y. Feng and M. Lapata, "Automatic Image Annotation UsingAuxiliary Text Information," Proc. 46th Ann. Meeting Assoc. of Computational Linguistics: Human Language Technologies, pp. 272-280, 2008.

[16] C. Buckley and E.M. Voorhees, "Retrieval System Evaluation," TREC: Experiment and Evaluation in Information Retrieval, E.M. Voorhees and D.K. Harman, eds., pp. 53-78, MIT Press, 2005.

[17] E.W. Noreen, Computer-Intensive Methods for Testing Hypotheses: AnIntroduction. John Wiley & Sons, Inc., 1989.

[18] D. Klein and C.D. Manning, "Accurate Unlexicalized Parsing," Proc. 41st Ann. Meeting Assoc. of Computational Linguistics, pp. 423-430, 2003.

[19] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "AStudy of Translation Edit Rate with Targeted Human Annotation,"Proc. Seventh Conf. Assoc. for Machine Translation in the Americas, pp. 223-231, 2006.

[20] A. Ahmed, E.P. Xing, W.W. Cohen, and R.F. Murphy, "Structured Correspondence Topic Models for Mining Captioned Figures in Biological Literature," Proc. ACM SIGKDD 15th Int'l Conf. Knowledge Discovery and Data Mining, pp. 39-48, 2009.

[21] R. Socher and L. Fei-Fei, "Connecting Modalities: Semi-Supervised Segmentation and Annotation of Images Using Unaligned Text Corpora," Proc. IEEE Conf. Computer Vision and PatternRecognition, pp. 966-973, 2010.

[22] J. Boyd-Graber and D. Blei, "Syntactic Topic Models," Proc. 22nd Conf. Advances in Neural Information Processing Systems, 2009.

[23] A. Sadeghi and A. Farhadi, "Recognition Using Visual Phrases,"Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1745-1752, 2011.

[24] D. Blei and M. Jordan, "Modeling Annotated Data," Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 127-134, 2003.

[25] D. Lowe, "Object Recognition from Local Scale-Invariant Features," Proc. IEEE Int'l Conf. Computer Vision, pp. 1150-1157, 1999.

[26] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," Int'l J. Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.

[27] M. kolajczyk and C. Schmid, "A Performance Evaluation of Local Descriptors," Proc. IEEE Conf. Computer Vision and Recognition, vol. 2, pp. 257-263, 2003.

[28] A. Bosch, "Image Classification for a Large Number of Object Categories," PhD dissertation, Universitat de Girona, Sept. 2007.

[29] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," J. Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[30] K. Sparck Jones, "Automatic Summarizing: Factors and Directions,"Advances in Automatic Text Summarization, I. Mani and M.T. Maybury, eds., pp. 1-33, MIT Press, 1999.

[31] I. Mani, Automatic Summarization. John Benjamins Publishing Co., 2001.

[32] G. Salton and M. McGill, Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[33] M. Steyvers and T. Griffiths, "Probabilistic Topic Models," A Handbook of Latent Semantic Analysis, T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, eds. Psychology Press, 2007.

[34] M. Witbrock and V. Mittal, "Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries," Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 315-316, 1999.

[35] R. Kneser, J. Peters, and D. Klakow, "Language Model Adaptation Using Dynamic Marginals," Proc. Fifth European Conf. Speech Comm. and Technology, vol. 4, pp. 1971-1974, 1997.