Big Data and Information Analytics
doi:10.3934/bdia.2017001
©American Institute of Mathematical Sciences

FIRST STEPS IN THE INVESTIGATION OF AUTOMATED TEXT ANNOTATION WITH PICTURES

J. Kent Poots and Nick Cercone

York University
Dept. of Electrical Engineering and Computer Science
4700 Keele Street, Toronto, Ontario, M3J 1P3, Canada

(Communicated by Jianhong Wu)

Abstract. We describe the investigation of automatic annotation of text with pictures, where knowledge extraction uses dependency parsing. Annotation of text with pictures, a form of knowledge visualization, can assist understanding. The problem statement is, given a corpus of images and a short passage of text, extract knowledge (or concepts), and then display that knowledge in pictures along with the text to help with understanding. A proposed solution framework includes a component to extract document concepts, a component to match document concepts with picture metadata, and a component to produce an amalgamated output of text and pictures. A proof-of-concept application based on the proposed framework provides encouraging results

1. **Introduction.** People view and interpret textual information to help perform activities of daily life. In situations constrained by time or capability, we sometimes overlook important information [5]. Today, technology is ubiquitous which augments our capabilities, in communication (the mobile phone), learning (the online course), and interpretation (the online encyclopedia). These technologies are universally applied because they clarify meaning and improve outcomes.

Some people have a cognitive deficiency and can't understand complex text, or can only process text with additional aid [8]. Others are "on-the-go" and busy, needing only a snippet here and a snippet there to address specific concerns, but they get overloaded by streams of text, which unhelpfully express simple concepts in complex ways. This work describes a means of providing a "cognitive augment" to help people understand concepts expressed in text, by adding pictures.

Psychologists say that pictures help us to understand text [3]. Studies in Human-Computer Interaction suggest many reasons for "providing assists" to text-only reading material: reading is unnatural, our attention is limited, and our memory is imperfect [5]. Recognition is easy, and recall is hard, which means that we understand pictures much faster than we interpret text [5].

Our approach to annotating text with pictures is novel because of a new combination of sub-components. We use dependency-analysis to extract concepts from text, and match those concepts to photo-realistic images; this combination is new

2010 *Mathematics Subject Classification.* Primary: 68T50; Secondary: 68T35.
Key words and phrases. Natural language processing, natural language understanding, information extraction, information visualization, artificial intelligence.
Financial support provided by York University is gratefully acknowledged.
* Corresponding author: Kent Poots *.

for the genre. Further, we look at the problem from two views, knowledge extraction and translation. This two-view approach is not seen in literature describing previous (related) systems.

1.1. **Objectives of this Work.** The objectives of this work are to test the feasibility of text annotation with pictures, to propose a framework (or pipeline) for performing this task, to test the framework, to refine the problem statement, and to identify relevant topics from Computational Linguistics (CL) for further research.

1.2. **Organization of presented material.** The material presented in this paper includes: Section 1: Introduction (this Section), Section 2: Background and Related Work, Section 3: The Proposed Framework and Test Results, and Section 4: Discussion, Conclusions, with Next Steps.

2. **Background and related work.** This section provides background information, including mentions of related prior work, in light of two possible approaches to the problem of automatically annotating text. The first approach views the problem as one of knowledge extraction and remapping. The second approach views the problem as one of translation, from a text language to a visual language.

The knowledge extraction view of this work is considered first, then we consider a view of the work as a translation.

Finally, we discuss some common cognitive science considerations termed rendering considerations.

2.1. **First view of the problem: Knowledge extraction and remapping.** Figure 1 shows a workflow or pipeline which extracts knowledge from input text, and provides a picture representation.

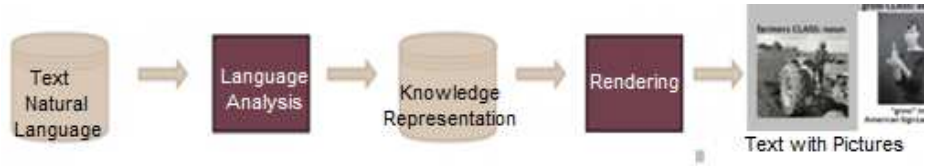

FIGURE 1. Processing for Text to Picture System

We note that, in the "Language Analysis" step, we can extract meaning at many levels, and this "Hierarchy of Meaning," or "Grammatical Rank" (see Figure 2) can become a parameter in the analysis (see [4]).

Components to extract knowledge (Figure 1) and transform that knowledge to an alternate form require a knowledge representation scheme. A form of logic can be used. McCarthy and other pioneers [5] were drawn to a logic representation because of the potential for well defined and unambiguous syntax, well defined semantics, and the potential for automated reasoning.

For example, to formally express a particular habit of cats, in a formal logic representation, we could write:

$\exists$x Sleeping(x,cat) .... which is the equivalent of *some cats are sleeping*

There are other, equivalent representations, including semantic nets, frames, rules, and scripts; ontologies can be used to fill gaps and resolve ambiguities [5].

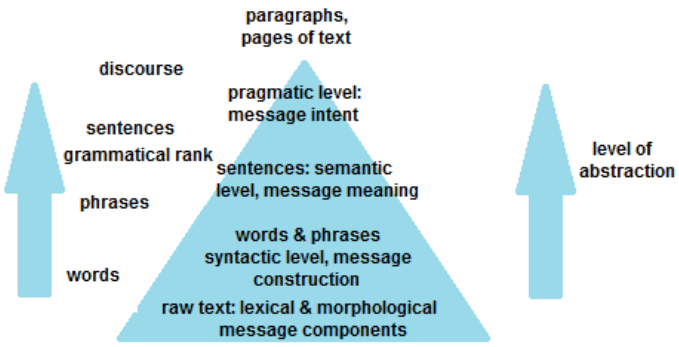

FIGURE 2. The Hierarchy of Meaning

We note that a vector representation of text concepts is possible (e.g., word2vec), but additional processing has only recently been formulated to adapt a vector form to current NLP knowledge bases (e.g., WordNet) and knowledge-focused applications (e.g., reasoning with knowledge). This is an area of current research.

We require knowledge extraction tools, to extract the facts, objects, and relations in text. There are least two approaches: statistically-based 'topic extraction" approaches, and linguistically-based "parsing," an example of which is syntactic dependency analysis.

The distinction between "statistically-based" (or "corpus-based" tools) and "linguistics-based" tools can be somewhat blurred - see tools described in the survey paper A comparison of extraction tools for the semantic web [1]).

2.2. **Linguistically-based approaches for knowledge extraction.** The few systems which attempt "text annotation with pictures" usually implement a linguistic-based approach. Dependency analysis (or parsing) is such a linguistic-based tool used in typical systems. We chose to use this form of analysis for our approach.

Dependency analysis relies on the premise that some words are "related-to" or "dependent" upon other words. Dependency is a binary asymmetric relation, that holds between a head word and its dependents [9]. This form of analysis can identify the "main action" and the "main subject" in a sentence.

The Stanford Dependency Parser [9] is usually considered the "gold standard" for extracting linguistic (syntactic) dependencies. Stanford form dependencies have been adopted for our work.

2.3. **Important prior work focused on concept extraction.** The most notable representative work is WordsEye [1], described by B. Coyne and R. Sproat in 2001. WordsEye produces 3D scene renderings using models from a text description. The Story Picturing Engine, another important contribution, is described by Joshi and Wang (2006) [4]. The focus is story illustration. Common system features of prior work include:

1. A processing pipeline.

2. Knowledge extraction tools such as a dependency parser.
3. External reference information to help resolve ambiguous terms.
4. Rendering, sometimes in 3D.

In the present research, we use this feature list as a starting point for the design of our system. More information about prior work is provided in the IEEE "in-press" paper *Automatic Annotation of Text with Pictures* [6].

2.4. **Summary the problem as knowledge extraction.** In this Section, we described natural language processing concepts which play a role in our solution to the problem of text annotation with pictures. Those components include knowledge extraction which requires a form of knowledge representation. There can be a hierarchy of meaning, and a hierarchy of processing for text, considering the meaning of words, sentences, paragraphs, or entire documents.

Briefly, we next review a second perspective of the text annotation with pictures problem, as a language translation problem. This view has proven useful in related work.

2.5. **Second view of the problem: Text language to visual language translation.** Translation is *the communication of the meaning of a source-language text using expression in an equivalent target-language text.* In our work, preservation of meaning is preferred over preservation of presentation. The objective is to provide communicative translation, which has a higher requirement for semantic accuracy, and a lesser requirement for structural likeness than other forms of translation.

Translation sets requirements on source language (SL) resources and the target language (TL) resources. Requirements include a vocabulary (lexicon), rules of construction (grammar), and rules for exceptions. There is an advantage in this work, in that the visual target language is as-yet unspecified. Language features can be selected (or an existing target visual language with those features can be adopted) to ease the translation task.

2.6. **Prior work focused on translation.** Prior work which "translates" text to visuals is mostly associated with translating text to sign language. The most relevant work is built on rule-based translation systems, described in the papers *A Machine Translation System from English to American Sign Language* [11] and *The Zardoz System* [10]. The sign language work is focused on a very specific (and helpful) purpose; it is not intended to draw from a pool of generalized images. More recent work includes the paper *Creating a High-Quality Translation System for a Low-Resource Language* [2], which implements "state of the art" phrase-based translation; this implementation requires parallel text.

Given a hypothetical visual language, with representations (and grammar rules) for nouns, verbs, sentences, and paragraphs, we could "map" text to this visual language, to translate from one language to another. For the problem at-hand, some of required data is missing. When visual language examples, grammars, and lexicons are more fully developed, we can pursue the translation approach; meanwhile, we will pursue the knowledge extraction approach.

2.7. **Visual rendering (cognitive) considerations.** Visual rendering is the process of transforming a representation, in our case a text representation, to a representation of an image or a scene. Visual representations can evoke a number of responses. Complex cognitive mechanisms are at work when we view a picture and try to make sense of that picture. Ideally, pictures would be created to optimize

interaction with these cognitive mechanisms. Any transformation of representation into the visual realm must consider the nature of human cognition, and what reaction a rendering might evoke.

2.8. **Summary the two views of the approach to a Solution.** In this Section, we considered two general approaches to the problem of text picturing, namely knowledge extraction and text to picture translation. Important prior work was discussed. We also discussed some visual rendering considerations derived from cognitive science. We mentioned notable prior work for both approaches. For the most part, existing systems favour linguistic-focused processing.

In the next section, we review the details of our Annotation Framework including test results.

3. **The annotation framework and test results.** Our proposed framework is an evolution of previous work, where we have addressed perceived gaps (see Figure 4). In this framework (or architecture), some components are optional in a basic implementation.

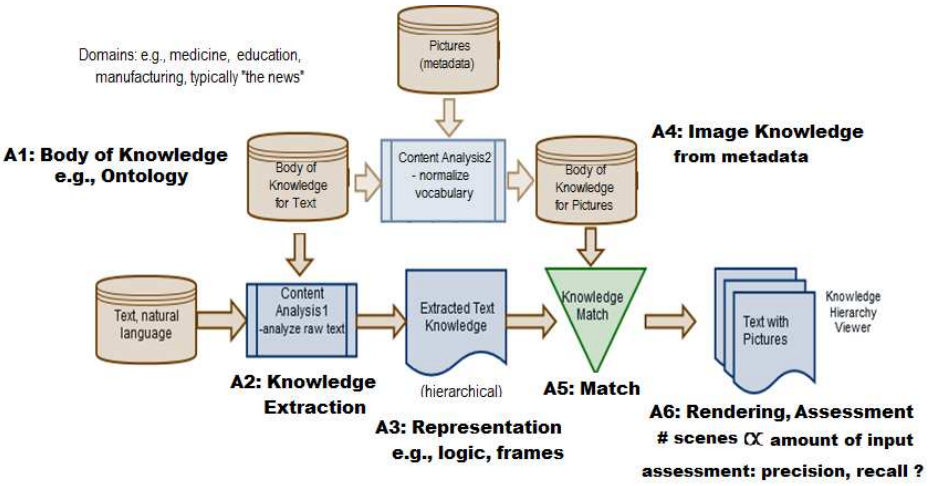

FIGURE 3. Proposed Annotation With Pictures Framework

Components of the framework include: external reference information to assist with knowledge extraction (i.e., "A1: Body of Knowledge"), knowledge extraction shown as "A2: Knowledge Parse," a means of representing the extracted knowledge labelled "A3: Representation," a picture database, and a means of matching text and pictures labelled "A4: Optimized Match".

External knowledge (A1) can be a formal ontology, and knowledge extraction (A2) could actually be an ontology-based extraction technique. Content analysis typically requires multiple passes, to extract named entities, and then extract the relationships associated with those entities. Image knowledge (A4) represents a database of (input text) compatible semantic information about images; processing is usually required to ensure compatibility. Matching (A5) can use a standard algorithm, such as cosine similarity.

1	Knowledge Representation	Subject / Verb / Object using Stanford form dependencies
2	Knowledge Extraction	RAKE baseline, Stanford, TextRazor, Deppattern parsers
3	Test Sentences	4 short sentences including French translation
4	Image Database	Google-Image pictures of nouns, verbs in Sign Language
5	Text/Image Matching	Binary match

TABLE 1. Core Components for Text Picturing Implementation

3.1. **Components used for core framework implementation.** Table 1 lists the core components implemented, including a means of knowledge representation and knowledge extraction, a small collection of test sentences, a baseline (called RAKE [7]) for knowledge (term) extraction, dependency parsers for knowledge (SVO) extraction, a rendering function, and integration code.

3.2. **Text picturing - core components - test plan.** The multi-modal (text/images) nature of the output presents challenges for evaluation. There are few standard (automated) metrics to assess the overall "goodness" of the result. In prior work, some teams asked users to "grade" the "understandability" of their systems output [1]-[4]. We choose to save user-testing for the next iteration of our framework, and instead we implemented testing on a pass/fail per-sub-component basis.

Our testing included the comparison of: dependency parser text output to a baseline RAKE output, the parsing output from multiple dependency parsers, subject-verb-object parser output to the concepts manually extracted from the input sentence, and the output image to the input text, judging if represented concepts actually match. English and French versions of test sentences were reviewed, to informally judge if output meaning would be understandable regardless of the caption language.

The following steps are undertaken for each test sentence:

1. Extract key terms (knowledge) using the RAKE function, to set a baseline for dependency analysis. Dependency analysis should not extract fewer terms than RAKE.
2. Generate a graphical constituency parse to evaluate sentence complexity (number of levels in the graph), and a dependency parse to retrieve SVO using each of the three extraction tools: Stanford, TextRazor, and Depparse. Compare those outputs to ensure consistency.
3. Compare the input sentence to the SVO output, checking if sentence concepts are represented in the output.
4. Process input text using the analysis pipeline to create a rendered scene. Evaluate if the rendered scene represents the sentence.
5. Identify gaps in knowledge extraction and rendering. If a gap can be addressed by a minor update, then address that gap, otherwise note the gap and move to the next sentence.
6. For a sample test sentence and a sample from "related work": compare input sentence complexity and output rendering coverage of input concepts.

7. For evaluating the utility of "picturing" as an aid to translation (e.g., English to French translation), evaluate the meaning of each output rendering without considering the input text.

3.3. **Test results and discussion.** This Section discusses results from the set of test tasks reviewed in the previous Section. We include qualitative observations on the tests previously described plus actual example output of a "pictured" sentence.

1	Extract baseline terms using RAKE	RAKE extracted meaningful terms
2	Constituency and dependency parse	Stanford and TextRazor gave same result; Depparse results varied
3	Compare input text to SVO	SVO was adequate for basic sentences
4	Create rendered scene	Scene matched SVO
5	Knowledge Extract, Rendering Gaps ?	SVO sometimes needs additional terms; consider verb valency.
6	Compare to prior work	Renderings provided equivalent detail.
7	Evaluate output pictures alone	Pictures could help understanding; pictures not a replacement.

TABLE 2. Core Components - Test Results Summary

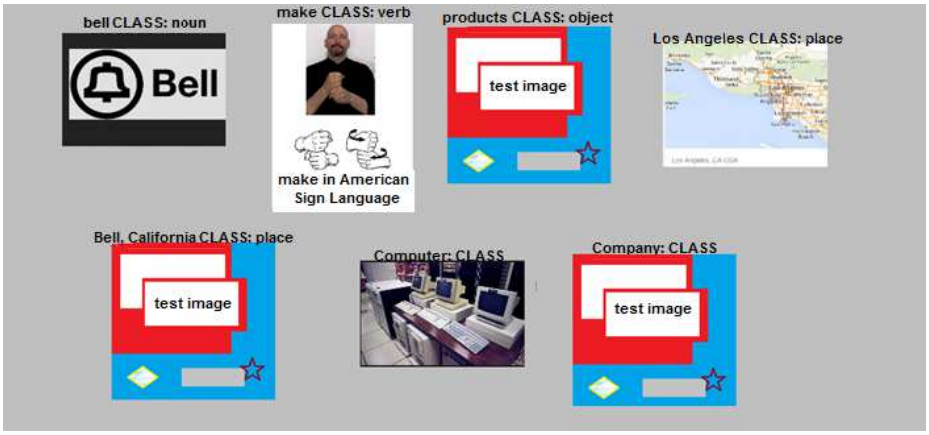

FIGURE 4. Picturing Results for Bell, a company which is based in Los Angeles, makes and distributes computer products

In the example of Figure 4, some concepts are shown as "generic test images," including products with CLASS object, "Bell, California" with CLASS place, and company. A "generic image" appears when a matching image cannot be found, in this case due to the picture collection small size. There is a representation of Bell making something in Los Angeles related to computers. This representation includes core concepts which are understandable even with some missing images.

A conclusion is that extraction of SVO plus an additional term , plus named-entities will provide reasonable results with simple sentences. For complex sentences, some concept features may be missed. A conclusion is that more complicated

sentences will require refinement to the simple SVO and named entity technique for concept extraction.

The sample result supports the findings of literature search, that some collapsing of representations would be helpful. In the "Bell" example, there are some terms which do not add much value to the representation. In addition to collapsing of dependency relations, a next iteration of the approach could see some weighting terms applied to the derived tags, and some checks for duplication: if a named entity appears in a derived relationship, it need not be represented more than once.

3.4. **Picturing as an aid to language translation.** We note that the meaning of each test sentence can be reasonably understood from the pictured result, irrespective of the language of annotation (both English and French are included for comparison).

3.5. **Comparison to a representative picturing system.** Figure 5 shows results from a contemporary "picturing" system. We note that the scene is composed of about 4 images, which identify the main nouns in the sentence. Actions (verbs) are not "pictured". A finding is that the results of our initial work are not immeasurably different from this contemporary system.

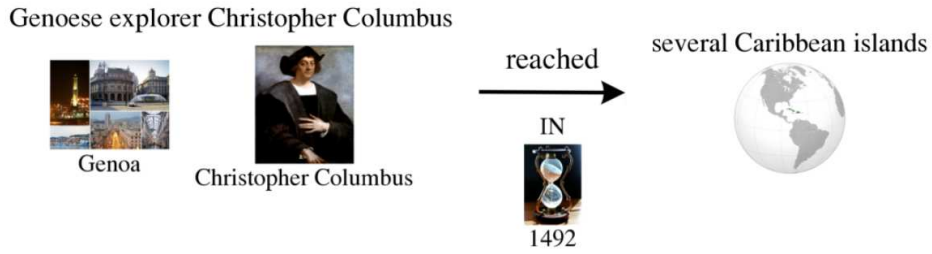

FIGURE 5. Example of Text Automatically Annotated With Pictures From UzZaman et al. [10]

3.6. **Summary of test results.** To summarize this Section about Results, we presented results about: the choice of knowledge extraction tools, the RAKE (baseline) results for keywords, graphical parse features of phrases of different complexity, and results about the usefulness of Subject / Verb / Object in representing the concepts in a sentence.

We present the finding that SVO is adequate for simple sentences, but can miss details for longer sentences. SVO is used by the representative picturing systems discussed as prior work.

Our system produces results which compare favourably to existing contemporary systems. Next iterations of our work will improve results, by employing additional features included in our suggested architecture.

The next section uses the findings of this section to determine if our work met research goals.

4. **Discussion, conclusion, next steps.** In this section, we review findings for this research in light of initial objectives. Table 3 provides a comparison. We also provide discussion about the Framework, refine the Problem Statement, and propose future research directions.

	Objective	Actual Result
1	Evaluate feasibility	Feasibility was demonstrated
2	Identify CL topic areas	Topics include IR, KE, parsing, matching, cognitive science (perception)
3	Propose a framework	Proposed, demonstrated dependency parsing, binary match
4	Provide Test Results	Demonstrated SVO model for short sentences; may need to consider term valence.

TABLE 3. Research Objectives vs. Actual Results

The refined Problem Statement is: *Provide an automated means of annotating text with pictures, considering available knowledge extraction methods, external knowledge sources (e.g., Wikipedia), and available picture database.*

4.1. **The proposed framework - discussion.** The proposed framework accepts text as input, extract concepts from that text using the (linguistic-based) approach of dependency parsing, and renders the input concepts as one or more images. We only touched-on issues of cognitive ability and how that might be involved in rendering.

If we include consideration of human cognitive ability, then the details of output visuals may change. At the root of a cognitive impact discussion are questions such as: *How do humans extract meaning from text ?* and *What representations provide the most insight ?*

Insights from cognitive psychology and the engineering field of Human Computer Interaction (HCI) about knowledge learning, knowledge representation, and knowledge retention could help. There is good overview discussion of human cognition in the overview work of Pinker *How the Mind Works*, the work of McGrath et al. *Visual learning for science and engineering.*

4.2. **Future research directions.** This investigation took steps toward demonstrating automatic annotation of text with pictures. There were some answers to initial research questions, and some findings. Many questions and topics were identified for followup.

The followup research questions include:

1. Was the method of knowledge extraction “optimal” ?,
2. Are there alternative approaches for knowledge extraction ?,
3. Can we consider statistical (or corpus) techniques ?
4. What are options for sourcing pictures ? and matching pictures ?

For many of the topics (and sub-topics) covered in this work, there is a significant margin for additional research. Automatic text annotation with text has been confirmed as an interesting, multi-faceted, challenging, interdisciplinary problem. [1]

REFERENCES

[1] B. Coyne and R. Sproat, WordsEye: An automatic text-to-scene conversion system, Proceedings of the 28th annual conference on Computer graphics and interactive techniques, **3** (2003), 487–496.

[2] D. Genzel, K. Macherey and J. Uszkoreit, Creating a high-quality machine translation system for a low-resource language: Yiddish, (2009), Available from: www.mt-archive.info/MTS-2009-Genzel.pdf

[3] A. Handler, An empirical study of semantic similarity in WordNet and Word2Vec, *Columbia University*, (2014).

[4] D. Joshi, J. Z. Wang and J. Li, The Story Picturing Engine—a system for automatic text illustration, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, **2** (2006), 68–89.

[5] J. McCarty, Programs with common sense, *Defense Technical Information Center*, (1963).

[6] J. K. Poots and E. Bagheri, Automatic annotation of text with pictures, (in-press), *IEEE IT Professional*, (2016).

[7] S. Rose, D. Engel, N. Cramer and W. Cowley, Automatic keyword extraction from individual documents, *Text Mining*, (2010), 1–20.

[8] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta and F. Ciravegna, Semantic annotation for knowledge management: Requirements and a survey of the state of the art, *Web Semantics: science, services and agents on the World Wide Web*, **4** (2006), 14–28.

[9] N. UzZaman, J. P. Bigham and J. F. Allen, Multimodal summarization of complex sentences, Proceedings of the 16th international conference on Intelligent user interfaces, **2** (2004), 43–52.

[10] T. Veale, A. Conway and B. Collins, The challenges of cross-modal translation: English-to-Sign-Language translation in the Zardoz system, *Machine Translation*, **13** (1998), 81–106.

[11] L. Zhao, K. Kipper, W. Schuler, C. Vogler, N. Badle and M. Palmer, A machine translation system from English to American Sign Language, *Envisioning Machine Translation in the Information Future*, (2000), 54–67.

E-mail address: kpoots@cse.yorku.ca
E-mail address: nick@cse.yorku.ca

[1] Last Word from the First Author: This paper incorporates some 2014 material where Professor Nick Cercone made an important contribution to the discussion of Natural Language Processing. It was a privilege to have had Professor Nick Cercone as my Ph.D. supervisor. Nick was a generous contributor to research.Thank you, Nick, for your wisdom, for your kind and able supervision, and for being a friend.