Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence

Cross Media Entity Extraction and Linkage for Chemical Documents

Su Yan, W. Scott Spangler, Ying Chen
IBM Almaden Research Lab
San Jose, California 95120
USA
{syan, spangles, yingchen}@us.ibm.com

Abstract

Text and images are two major sources of information in scientific literature. Information from these two media typically reinforce and complement each other, thus simplifying the process for human to extract and comprehend information. However, machines cannot create the links or have the semantic understanding between images and text. We propose to integrate text analysis and image processing techniques to bridge the gap between the two media, and discover knowledge from the combined information sources, which would be otherwise lost by traditional single-media based mining systems. The focus is on the chemical entity extraction task because images are well known to add value to the textual content in chemical literature. Annotation of US chemical patent documents demonstrates the effectiveness of our proposal.

Introduction

Images have always been a major source of information in scholarly articles, such as journals, proceedings, patents, technical reports and books. Along with text, images convey key concepts and major contribution of an article. From the standpoint of automated document understanding and knowledge extraction, an important but little-studied part of scholarly articles is the connection between text and images. Textual content and images within an article are not independent, rather the two media reinforce and complement each other. Individually mining each media can only provide partial information of the entire document.

Existing data analysis and information extraction (IE) techniques are usually designed to target at a particular media type and are not applicable to data generated by a different media type. For example, existing entity extraction techniques focus on textual data. Entities of interest, such as protein and gene names (Krauthammer et al. 2000), chemical names and formulae (Sun et al. 2007; Klinger et al. 2008), drug names (Hamon and Grabar 2010) etc., are automatically extracted from the textual part of a document. The important information conveyed by images is discarded and made inaccessible to users. Moreover, several image analysis and computer vision techniques are developed to automatically extract visual components of interest from article images, such as caption extraction and indexing (Cohen, Wang, and Murphy 2003), diagram summarization (Lu et al. 2007; Futrelle 2004), chemical structure extraction (Park et al. 2009) etc.. These techniques do not consider the textual context where the images appear, thus the connection between images and text is neglected.

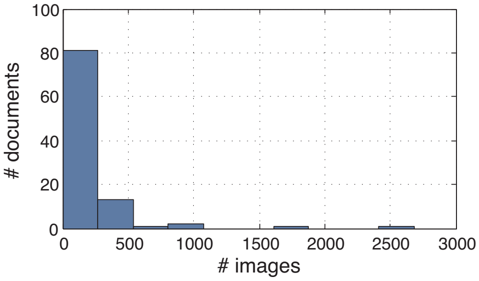


Figure 1: Histogram of number of images per chemical document (mode = 137)

Simply combining text-based and image-based information extraction techniques will not solve the problem, since the semantic links exist between the two media are not explored by either technique. In this work, we propose an IE scheme that explores the structural and language characteristics of chemical documents to bridge the gap between the visual content represented by images and the textual content represented by words. The scheme integrates text analysis and image processing techniques to jointly mine the two media and is able to discover the knowledge which is otherwise lost by traditional single-media based mining systems.

Images are particularly important for chemical literature, because the 2D depiction of a chemical structure is the preferred representation of chemicals in the community (Zimmermann and Hofmann-Apitius 2007). For this reason, key entities in chemical articles are all introduced in images. Figure 1 shows the histogram of the number of images per document for 100 randomly selected US chemical patents. Some documents contain more than 2K images each. Of all the images, 99.01% are depictions of chemical structures. The mode is 137 structure-depiction images per document.

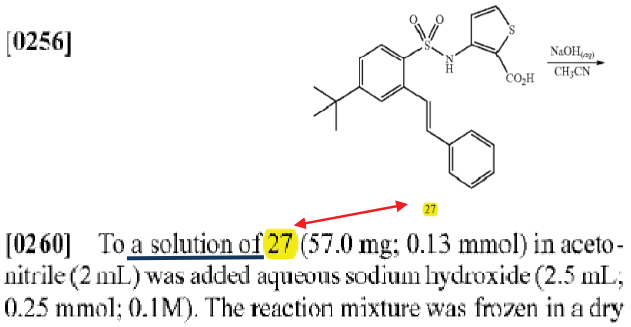
Recently, there is extensive interest in automatic chemical entity extraction from documents. Unlike other entity types such as people names, addresses, phone numbers etc., chemical entities are typically defined both in text and in images. Independently mining text or images cannot reveal all the information. Moreover, a recognized yet difficult to solve problem in chemical document analysis is *anaphora* (Banville 2006). A chemical is usually defined with an assigned *chemical label* when first being introduced in a document. When the chemical is referenced in the later part of the document, the label, instead of the actual chemical name or structure, is used. For example, Figure 2 shows an extraction from a patent document. A chemical structure with label "27" is introduced in claim [0256] and is referenced in claim [0260] by using the assigned label to describe an experimental procedure. Existing chemical entity extraction solutions do not handle anaphora. Imagine a search service is built upon extracted chemicals. If a user queries a chemical and expects to get a list of context snippets to understand how the chemical appears and is used in documents, a large amount of information is inaccessible due to anaphora. Anaphora also harms other services, such as ranking chemicals by importance or relevancy, similarity comparison of chemical documents, or chemical entity relation identification. Our work handles the anaphora problem and makes the hidden information available to end users.

A *chemical label*, which is the unique identifier of a chemical structure is typically used in chemical documents to link images to text and associate the textual content of the entire document. For the chemical entity extraction task, the main challenge is to extract and associate chemical structures to corresponding labels, as well as associate mentions of labels. Based on the visual structure of images, we automatically extract chemical structures and labels, and map every extracted structure to the corresponding label. The mapping pairs are stored. We further analyze the textual part of a document to locate all the mentions of chemical labels. The difficulty of chemical label extraction from text is due to the fact that a label can appear in text in different scenarios. For example, the label text "5" can appear in "compound 5", "Table 5", "experiment 5" or in chemical name "bis (4 hydroxy 5 nitrophenyl)methane" [1]. We introduce *context-aware CRF* (caCRF), an efficient information extraction method to explore the context of the label text, and define novel features to extract chemical labels only. The labels extracted from the two media are then used to link images and text.

In summary, we present an integrated scheme to collaboratively extract and link chemical entity information found in text and images. Our goal is to provide an automatic document analysis solution that incorporates information from various information sources.

Related Work

Previous work in the area of information extraction has largely focused on single-media based extraction. For example, techniques for named entities extraction or entity relations extraction are typically designed for textual data. Recent activities in multimedia document processing, such as concept extraction from images/video/audio (Yan and Naphade 2005) also target at a particular media type. Our work differs from existing work in collaboratively extract and link information cross media.



[0256]

[0260] To a solution of 27 (57.0 mg; 0.13 mmol) in acetonitrile (2 mL) was added aqueous sodium hydroxide (2.5 mL; 0.25 mmol; 0.1M). The reaction mixture was frozen in a dry

Figure 2: Example of anaphora and label context

Previous work on chemical entity extractions falls into two categories: 1. those extract entities from text (Sun et al. 2007; Klinger et al. 2008; Chen et al. 2009), and 2. those extract entities from images (Filippov and Nicklaus 2009; Park et al. 2009). Entities from the same document that are extracted by the above two types of techniques share no semantic links. Our work is motivated by this observation and aims to link different media sources for more comprehensive information extraction. Note that, there is existing work that leverages the connection between text and images to identify useful information (Deserno, Antani, and Long 2009; Hua and Tian 2009). For example, text-based image indexing and searching techniques rely on image captions or surrounding text to generate image meta-words. These services use text content to boost the understanding of images, but do not set up explicit links between text and images to improve the overall analysis of a document.

System Overview

The overall workflow is outlined in Figure 3. Given an input document (PDF, Doc, HTML etc.) which contains text and images, a typical preprocessing step is to extract images and partition the original document into two parts: the plain text and the images. For example, our data set contains patent documents in the PDF format. In order to analyze these data, a PDF file is passed through an Optical Character Recognition (OCR) process to extract images and generate plain text. Advanced OCR solutions are able to preserve the location of extracted images so that analysis of image context is possible. We analyze images to extract chemical structures and labels. The extracted structures are fed into a professional chemistry software Cambridge Name=Struct® (Brecher 1999) to generate equivalent string representation which are called SMILES strings. SMILES are useful for text-based storage, indexing, and search. We further map each extracted structure to its corresponding label, and store

[1] The chemical name is extracted from real data, and has OCR errors. Correct name should be "bis(4-hydroxy-5-nitrophenyl)methane"

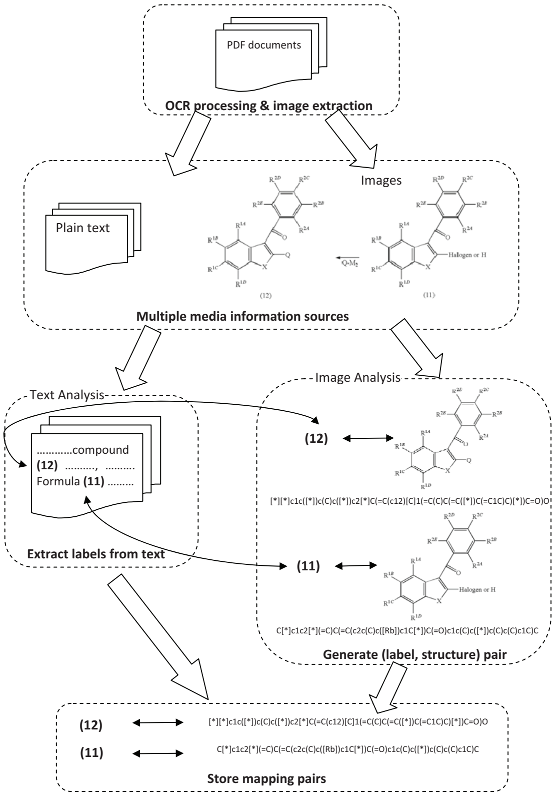

PDF documents
OCR processing & image extraction
Plain text
Images
Q-M
(12)
(11)
Halogen or H
Multiple media information sources
Text Analysis
............compound
(12) ...., ...
Formula (11) ....
Extract labels from text
Image Analysis
(12)
(11)
[*][*]c1c([*])c(C)c([*])c2[*]C(=C(c12)[C]1(=C(C)C(=C([*])C(=C1C)C)[*])C=O)O
C[*]c1c2[*](=C)C(=C(c2c(C)c([Rb])c1C[*])C(=O)c1c(C)c([*])c(C)c(C)c1C)C
Generate (label, structure) pair
(12)
[*][*]c1c([*])c(C)c([*])c2[*]C(=C(c12)[C]1(=C(C)C(=C([*])C(=C1C)C)[*])C=O)O
(11)
C[*]c1c2[*](=C)C(=C(c2c(C)c([Rb])c1C[*])C(=O)c1c(C)c([*])c(C)c(C)c1C)C
Store mapping pairs

Figure 3: Overall workflow

the (label, structure) pairs. Meanwhile for plain text, we perform statistical information extraction techniques to extract mentions of chemical labels. Given image and text extractions, labels serve as a connection of the two media. The anaphora problem is solved since chemical entities as well as their references are all identified and extracted.

Extract (Label, Structure) Pairs from Image

Compared to the images from other domain, chemical images have the following specialties: 1. An image typically does not contain caption or indicating key words such as “Figure” or “Fig.”. Therefore images cannot be linked to text by using captions or index numbers. 2. Labels to chemical structures are commonly defined in images. 3. An image usually contains multiple structures and labels. 4. Image layouts are complex. An image can depict a chemical reaction process which involves multiple steps and chemicals.

We introduce an efficient yet robust solution to analyze chemical images. Given an input image[2], we normalize the image by converting the RGB representation to grayscale representation and binarize the image. We further segment the image, and categorize the segments into two groups: those containing structures and those containing labels. A graph-matching algorithm is then applied to map structures to labels. The extracted (structure, label) pairs are stored for downstream processing.

Image Segmentation

Morphology-based image segmentation techniques are used to segment a chemical image into disconnected components. Given a chemical image, we first detect edges in the image, then dilate [3] the image in two steps.

Because chemical structures are basically made of lines, two flat and line shaped structuring elements are defined with 0° and 90° respectively. The horizontal element fills gaps in textual part of the image since horizontal writing is typically used in chemical images. The vertical element is useful to connect atoms to bonds. The image is further dilated with a disk-shape structuring element. The second step dilation fills tiny gaps that are missed by the vertical and horizontal elements. We also fill the holes in the image to ensure that no embedded segment is generated. An original chemical image and its dilation and filling result are shown in Figure 4 (a) and Figure 4 (b) respectively.

The dilated image contains several disconnected components. We enclose each component with a *minimum rectangle*. To do this, a background image (every pixel is of digit value 0) with the same size of the image is created. For each component, we copy the component to the background image at the same location as it appears in the original image. The background image can be treated as a matrix containing 0s (background) and 1s(foreground). To find the vertical boundaries of the component, we scan each column of the matrix from left to right. The first and last encountered columns with 1s are the left and right boundaries respectively. Similar scanning scheme is used to find the top and bottom boundaries. The minimum rectangles should not intersect with each other, otherwise, the two rectangles will be merged and a larger rectangle is generated to contain components from both intersecting rectangles. Figure 4 (b) shows an example of minimum rectangles.

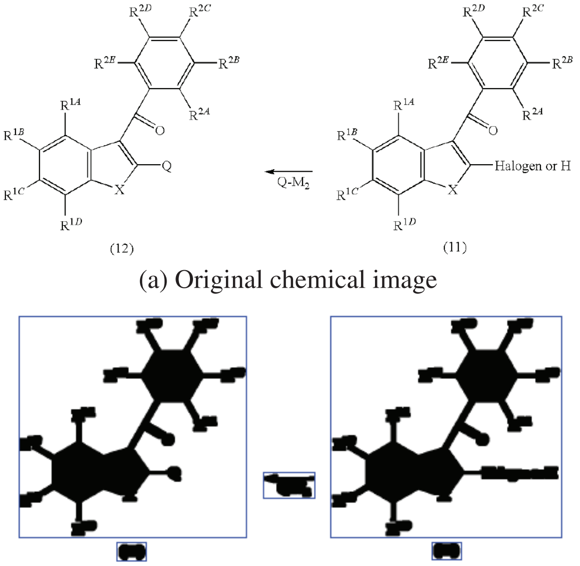
The rectangle enclosed component is then deleted from the original image and saved as an individual image, which we call it a *segment*. At this point, we generated a set of segments from the image. For example, 5 segments will be generated from the dilated image shown in Figure4 (b), and each segment if contains chemical structures, will contain one and only one structure.

Segments Categorization

To enable text-based storage, indexing and search, we leverage a professional chemistry software OSRA (Filippov and Nicklaus 2009) to convert 2D depiction of chemical structures to equivalent string representations. The conversion is computationally expensive. Our goal is then to efficiently identify chemical segments, which are segments that contain structures, and only feed the selected segments to OSRA for more complicated image processing. To this end, we pro-

[2] Our images are of 300 dpi. All the parameter settings reported in this work are based on this resolution.

[3] Dilation is a fundamental morphology operation that causes an object to grow in size in a controlled manner as defined by the *structuring elements*.



(a) Original chemical image

(b) Dilated chemical image with minimum rectangles

Figure 4: Morphology-based image segmentation

pose to use the following features that are cheap to generate and measure:

Segment Density: Compared to non-chemical segments, a large portion of a chemical segment is white space. Visually, a non-chemical segment is more "dense" than a chemical segment. We define *segment density* (SD) as the percentage of foreground pixels (e.g.: pixels with digit value 1) in the segment. Let l and w be the length and width of the minimum rectangle of a segment respectively, with $w < l$, segment density is:

$$SD = \frac{sum[pixel = 1]}{l \times w}$$

We consider segments with $SD > \alpha$ as non-chemical segments, where α is a picked threshold.

Segment Size: The 2D depiction of a chemical structure takes certain space. A too small segment is unlikely to contain structure depictions. Therefore the *size of a segment* (SS) is a good indicator. We consider segments with $SS = w < \beta$ as non-chemical segments, where β is a picked threshold.

Aspect Ratio: The aspect ratio of a chemical segment should be within a range. For example, a segment with the shape of a narrow strip is unlikely to contain chemical structures. Therefore we define *aspect ratio* (AR) of a segment as $AR = w/l$ and consider anything with $AR < \gamma$ as non-chemical, where γ is a picked threshold.

Thresholds α, β, γ are tuned using training data.

Label Segment Identification

As Figure 4 shows, a non-chemical segment does not necessarily contain a label, but can contain formulae, description of reactions etc.. In order to identify labels, we feed every non-chemical segment to an OCR engine[4] to extract text. We handle two cases: 1. If the segment contains a single token, we check if the token matches at least one of the features listed in Table 1. The segment is considered to contain a label if at least one match is found; 2. When the segment contains multiple tokens, if a label is defined, the label is defined within a context. In such cases, an indicating keyword, such as "compound", "formula" etc. is often used prior the label. We extract such keywords that are used in chemical documents and list them in Table 1. To identify the occurrence of a keyword, due to possible OCR errors, strict string matching does not perform well. For each token, we measure the Levenshtein distance (Wagner and Fischer 1974) to every keyword. If the distance is within an allowed range, the token is considered as an indicating keyword, and the following token is checked using the step 1. measure.

Table 1: Chemical Label Features

feature	example
consecutive Roman digits	II, XIV, iv
capital letters	B, IIA
pure digits	9
combination of digits and letters	D2, J3
letters connected by a dash	II-iia
digits connected by as dash	12-3
above features contains prime	IV'
above features enclosed by parenthesis	(XIV), (12-3), (D2)
keywords	
compound, formula, structure, preparation, example	

After labels are extracted, we also define OCR correction rules based on the error statistics on the training set, e.g. "VH" to "VII", "(D" to "(I)", "21" to "a" etc.

Link Label to Structure with Graph Matching

At this step, we have a list of a chemical segments denoted as $S = \{s_1, \ldots, s_a\}$, and a list of b label segments denoted as $L = \{l_1, \ldots, l_b\}$. a and b do not necessarily equal. For example, some structures are drawn for one time illustration and do not come with labels for later references in the text. We define the label-structure-mapping task as a *minimum-weight graph matching* problem. In particular, we define a bipartite graph $G = (S \cup L, E)$, $|S| = a$, $|L| = b$, where S is the vertex set of structures, L is the vertex set of labels, and E is the set of edges between S and L. We measure all the pairwise distances between the two sets of segments so G is complete. To evaluate edge weights, we treat the left bottom of an image as the origin, and measure the centroid location for each segment accordingly. The edge weight $\omega(e_{i,j})$ between a structure segment s_i and a label segment l_j is then defined as the Euclidean distance between the centroids of the two segments:

$$\omega(e_{i,j}) = dist(l_i, s_j) = dist(Z(l_i), Z(s_j))$$

where $Z(l_i) = (x_i, y_i)$ represents the x and y coordinates of the centroid of label segment l_i, and similar for $Z(s_j)$. Our goal is then to find a matching M, where $M \in E$ and no

[4] We use tesseract-ocr, an open source OCR tool

two edges in M share an endpoint, such that the total weight of M is minimized:

$$\min \omega(M) = \min \sum_{e \in M} \omega(e)$$

For a bipartite graph, the Dijkstra algorithm can be used to solve the problem in $O(n(n \log n + m))$ time, with n vertices and m edges.

Extract Labels from Text

So far, we have a list of chemical labels extracted from images. In order to extract chemical labels from text, an intuitive solution is to perform string matching and extract all the occurrences of a label. As we explained before, a label can appear in text in several different scenarios so this solution will generate too many false positives. Another solution is to perform rule-based extraction as we did for extracting labels from images. Text in images is simple so rule-based extraction works well. However, in the document body, due to the complexity of text, it is difficult to define precise and comprehensive set of rules to extract all the chemical labels.

Context-Aware CRF (caCRF)

Based on the above observation, we introduce a *context-aware Conditional Random Field* (caCRF) method to extract chemical labels from text. CRF (Lafferty, McCallum, and Pereira 2001) is the state-of-the-art sequence labeling and information extraction technique. Traditional CRF methods scan the entire document token by token and label each token to predefined categories. The computational cost is large for long documents. We propose to efficiently identify a small subset of a document and only apply CRF to the subset.

First of all, we define a *sequence* $T = \{t_i, \ldots, t_n\}$ as an ordered list on tokens $t_i, \ldots, t_n$ where the order is as they appear in a document. For the label extraction task, we define simple rules to quickly extract a list of "label candidates" $LC = \{g_1, \ldots, g_r\}$ from text. The rules are general enough to cover all the labels but with false positives. Then for each candidate g_i, a "context" $C(g_i) = \{T_i^b, g_i, T_i^a\}$ is extracted, where T_i^b and T_i^a are sequences before and after the label candidate. Therefore, a "context" is a sequence too. Given two contexts $C(g_i)$ and $C(g_j)$, if T_i^a and T_j^b overlap, we merge the two context to generate a new context that includes both candidate labels $C(g_i, g_j) = \{T_i^b, g_i, T_{ij}, g_j, T_j^a\}$. We fix the length of sequence T_i^b and T_i^a to be 5 tokens unless a sentence boundary is met.

Textual Feature Set

Besides the common features used in CRF methods, such as all the tokens, we extract two types of features. *Structural features*, such as token length and the features listed in Table 1, capture the composition of a token. *Content features* are generated based on the observation that chemical labels are often (not always) referenced by using indicating keywords (as listed in Table 1) and key phrases, such as "hereinafter referred to as (IV)", "a solution of 27" (Figure 2). One content feature is defined to indicate whether a token is an indicating keyword or a part of a key phrase. Other content features include whether the focus token is before or after the nearest keyword/key phrase, and the distance from the focus token to the nearest keyword/key phrase in terms of the number of tokens in between.

Table 2: Parameter Setting

parameter	value
horizontal structuring element size	9 pixels
vertical structuring element size	5 pixels
disc structuring element size	1 pixel
Segment Density α	0.09
Segment Size β	25 pixels
Aspect Ratio γ	0.2

Experiment

We evaluate our proposal using 100 chemical related US patent documents published in year 2010. To identify if a document is chemical-related, we apply IBM SIMPLE chemical annotator (Chen et al. 2009) to each 2010 US patent and select those that contain at least 200 unique chemical names to compose our dataset. The 100 documents contain 18,766 images, where 18,581 images contain chemical structures, and 2,882 images contain both structures and labels (many images contain structures only), which is about 15%. We perform 5-fold cross validation, and report the average result. The best parameter settings are reported in Table 2. Classic evaluation metrics *Accuracy*, *Precision*, *Recall* and *F score* are adopted.

The performance of image analysis and (label, structure) pair extraction is listed in Table 3. The image segmentation method has high accuracy with a few errors of missing atoms from the main structure. The segmentation accuracy can be improved by using advanced visual features and chemical domain knowledge. We achieve perfect segment categorization accuracy. For the label identification task, since we are using rule-based method, the identification accuracy can be improved by refining rules. For the pair extraction task, because many images contain more than one chemical structure, we consider an extraction method performs a correct extraction on an image if all the (label, structure) pairs from that image are correctly extracted and no extra noise is extracted. We evaluated three cases. The "overall" case measures extraction precision and recall for all the chemical structure images. For the "easy" case, we measure extraction performance on images that contain a single structure only. Since the layout of such images are relatively simple, we call such case "easy". The "difficult" case measures extraction performance on images that contain more than 5 chemical structures, which leads to more complicated layouts. As Table 3 indicates, the overall extraction performance is promising. We achieve around 90% of extraction accuracy and recall. When the image layouts become more complicated, the extraction accuracy drops as can be expected.

The text analysis and extraction performance is reported in Table 4. In the "exact" method, given a list of chemical

Table 3: Image Analysis Performance

operation	accuracy
image segmentation	98.12%
segment categorization	100%
label identification	96.97%

(label, structure) pair extraction

method	precision	recall	F score
overall	89.04%	93.88%	91.40%
easy	97.69%	100%	98.83%
difficult	77.55%	93.65%	84.84%

Table 4: Text Analysis Performance

method	precision	recall	F score
exact	47.56%	41.55%	44.35%
rule-based	28.66%	94.12%	43.95%
caCRF	90.91%	82.19%	86.33 %
CRF	90.91%	82.19%	86.33 %

labels extracted from images, we do strict string matching to extract all the label appearances from text. The low precision is because the label text can appear in many scenarios other than indicating a chemical entity. Extracting irrelevant appearance harms precision. Moreover, a label can be mentioned in text with a slightly different format as introduced in images. For example, the label "(IIX)" can be referred to as "IIX", "(IIX);", "IIX," etc. For this reason, strict string matching will miss many label appearances and has low recall. In the rule-based method, we specify rules about the composition of a label, similar to what we did in image analysis. As can be expected, this method has high recall, but generates many false positive and has low precision. The caCRF method achieves reasonable extraction performance. Moreover, the scheme of pre context selection significantly reduces the amount of data to be processed by CRF without influencing extraction accuracy. The amount of reduction is measured in terms of the number of tokens to be labeled by CRF, and we achieve 66.81% of reduction.

Conclusion

In this work, we propose an IE scheme that explores the structural and language characteristics of chemical documents to bridge the gap between the visual content represented by images and the textual content represented by words. The scheme jointly mines the two media and is able to discover the knowledge which is otherwise lost by traditional single-media based mining systems.

References

Banville, D. 2006. Mining chemical structural information from the drug literature. *Drug Discovery Today* 11(1-2):35–42.

Brecher, J. 1999. Name=struct: A practical approach to the sorry state of real-life chemical nomenclature. *Journal of Chemical Information and Computer Science* 39(6):943–950.

Chen, Y.; Spangler, S.; Kreulen, J.; and etc. 2009. Simple: A strategic information mining playform for ip excellence. *IBM Research Report*.

Cohen, W. W.; Wang, R.; and Murphy, R. F. 2003. Understanding captions in biomedical publications. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, 499–504.

Deserno, T. M.; Antani, S.; and Long, L. R. 2009. Content-based image retrieval for scientific literature access. *Methods of Information in Medicine* 48(4):371–80.

Filippov, I. V., and Nicklaus, M. C. 2009. Optical structure recognition software to recover chemical information: Osra, an open source solution. *Journal of Chemical Information and Modeling* 49(3):740–743.

Futrelle, R. P. 2004. Handling figures in document summarization. In *Text Summarization Branches Out Workshop, 42nd Annual Meeting of the Association for Computational Linguistics*, 61–65.

Hamon, T., and Grabar, N. 2010. Linguistic approach for identification of medication names and related information in clinical narratives. *Journal of the American Medical Informatics Association* 17(5):549–554.

Hua, G., and Tian, Q. 2009. What can visual content analysis do for text based image search? In *IEEE international conference on Multimedia and Expo*, 1480–1483.

Klinger, R.; Kolářik, C.; Fluck, J.; Hofmann-Apitius, M.; and Friedrich, C. M. 2008. Detection of iupac and iupac-like chemical names. *Bioinformatics* 24:i268–i276.

Krauthammer, M.; Rzhetsky, A.; Morozov, P.; and Friedman, C. 2000. Using BLAST for identifying gene and protein names in journal articles. *Gene* 259(1-2):245–252.

Lafferty, J. D.; McCallum, A.; and Pereira, F. C. N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, 282–289.

Lu, X.; Wang, J. Z.; Mitra, P.; and Giles, C. L. 2007. Deriving knowledge from figures for digital libraries. In *International conference on World Wide Web*, 1229–1230.

Park, J.; Rosania, G. R.; Shedden, K. A.; Nguyen, M.; Lyu, N.; and Saitou, K. 2009. Automated extraction of chemical structure information from digital raster images. *Chemistry Central journal* 3(1).

Sun, B.; Tan, Q.; Mitra, P.; and Giles, C. L. 2007. Extraction and search of chemical formulae in text documents on the web. In *International conference on World Wide Web*, 251–260.

Wagner, R. A., and Fischer, M. J. 1974. The String-to-String Correction Problem. *Journal of the ACM* 21(1):168–173.

Yan, R., and Naphade, M. 2005. Multi-modal video concept extraction using co-training. *Multimedia and Expo, IEEE International Conference on*.

Zimmermann, M., and Hofmann-Apitius, M. 2007. Automated extraction of chemical information from chemical structure depictions. *Drug Discovery* 12–15.