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* $T = \{t_i, \dots, t_n\}$ as an ordered list on tokens $t_i, \dots, t_n$ 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" $LC = \{g_1, \dots, g_r\}$ 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	2 pixels
vertical structuring element size	5 pixels
disc structuring element size	1 pixel
Segment Density δ	0.09
Segment Size ζ	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