Automated Generation of Ground Truth Data of Chemical Equations from Document Images

Prerana Jana, Anubhab Majumdar, Sekhar Mandal
Department of Computer Science and Technology
Indian Institute of Engineering Science and Technology Shibpur, India
Email: (prerana.jana, anubhabmajumdar93)@gmail.com
sekhar@cs.becs.ac.in

Bhabatosh Chanda
Electronics and Communication Sciences Unit
Indian Statistical Institute,Kolkata, India
E-mail : chanda@isical.ac.in

***Abstract*—The abstract goes here.**

I. Introduction

Here goes the introduction and previous work

II. Proposed Work

Major steps involved in the automated generation of ground truth data are given below.

A. Segmentation of displayed chemical equation
B. Recognition of various chemical symbols present in chemical equations
C. Optical character recognition of each reactant
D. Auto correction of reactants and products in chemical equations
E. Generation of ground truth data in PDF format

The details of the above steps are given in the following subsections.

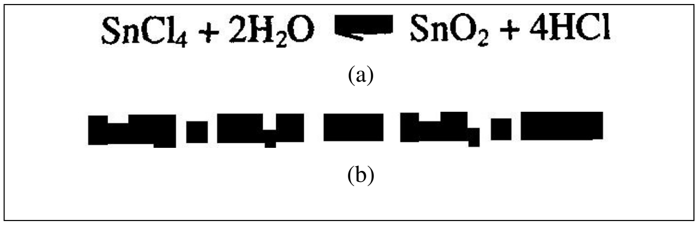
A. Segmentation of displayed chemical equation

B. Recognition of various chemical symbols present in chemical equations

A chemical equation is a way of representing a chemical reaction in symbolic form. Chemical equations consists of reactants separated by myriad chemical symbols. The symbols along with their significance are listed below.

- + : Separate the reactants
- →, ← : Separate the reactants from products in irreversible reactions; also denote the direction of reaction
- ↔, ⇌ : Separate the reactants from products in reversible reactions
- = : Shows stoichiometric equality in chemical equations
- ↑ : Used to denote gaseous compound
- ↓ : Used to denote sediments formed after a reaction

We begin with the extracted displayed chemical equation(DCE) from the step above and run a HRLS algorithm. This results in the coalescing of the chemical compounds into a word blob as shown in Fig.1b.



(a)

(b)

Fig. 1: (a). Original DCE. (b). DCE after closing operation.

All the single components are identified and segregated from the original equation and classified using a decision tree shown in Fig.2.

The function of each node of the decision tree is detailed below.

- #of component : This module count number of disjoint components in the input symbol.
- Crossing : We measures the number of transitions from object to background pixel or vice versa for each column while moving along the rows and consider the maximum value.
- OCR : The input symbol is run through OCR to identify positively the + symbol. Other symbols return erronous result.
- Aspect Ratio : Calculates the $\frac{height}{width}$ ratio of input image.
- Distance Transform :

Certain chemical element like Carbon, Sulfur etc.(see Fig.3) may also be input to the tree and they get erroneously identified as up or down arrow because characters have AR>0.8.

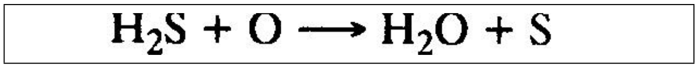


Fig. 3: DCE having element represented by single character

To fix this mess, we match the symbols classified with the original equation blob image. For each +, →, ←, ↔, ⇌ and = in the classified symbol set, we check its immediate right

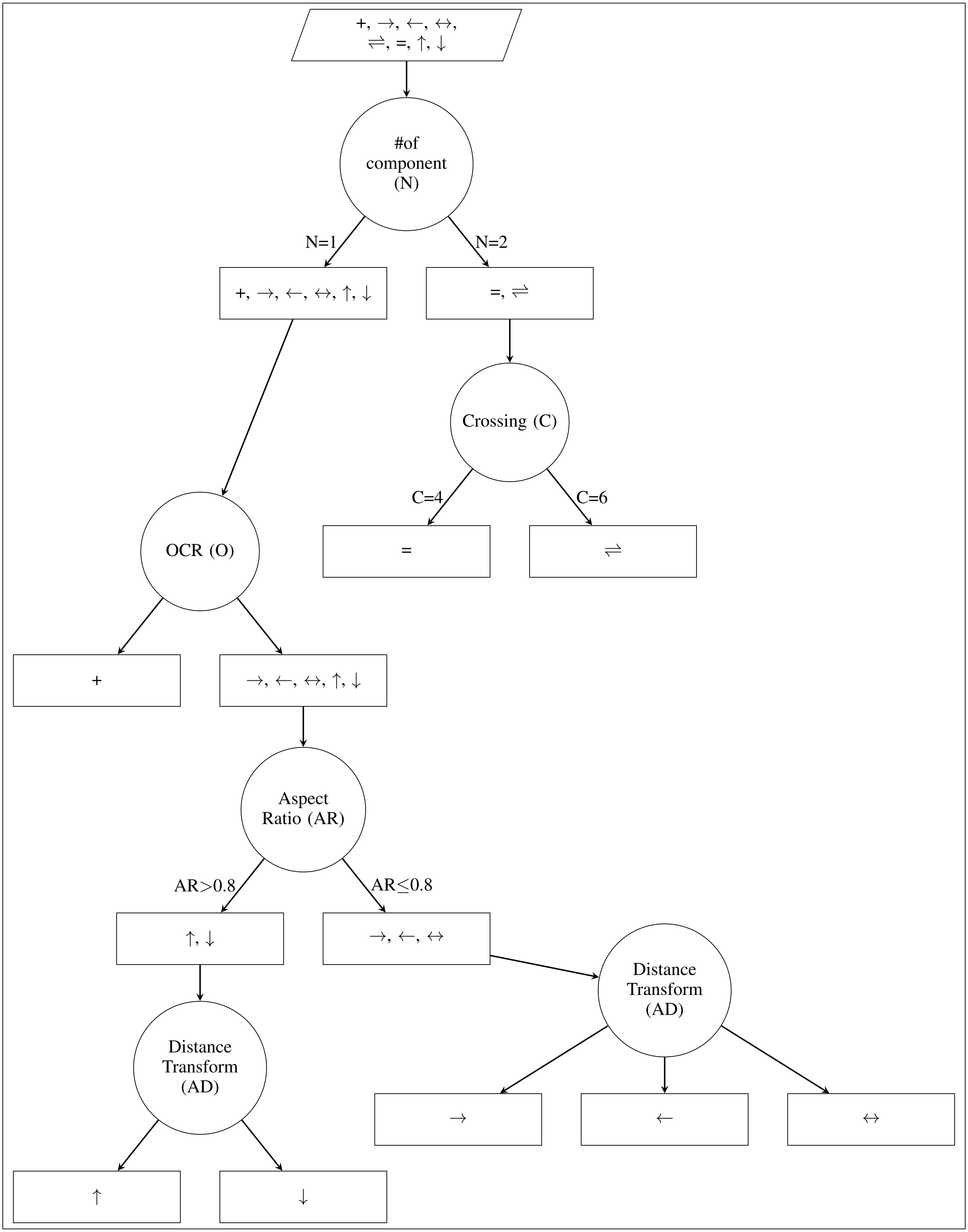


Fig. 2: Decision Tree for blaah Chemical Symbol Classification

blob. If the blob has single component in the original equation image, then it must be an single character element because these symbols must be followed by reactants. Also the first blob is checked to see if its a single character element or not. Thus the chemical symbols are successfully segregated and classified from the chemical equation.

C. Optical character recognition of each reactant

Each segment of a displayed chemical equation is divided into three zones; namely upper zone, middle zone and lower zone (see Fig ****). To identify the three zones of a DCE zone, uppermost and lowermost co-ordinates of each connected component below the same DCE zone are also obtained. The median of uppermost coordinate, and median of lowermost co-ordinate of such components in DCE zone are computed. A horizontal line, called the baseline, is drawn through the median of lowermost coordinates of components and this baseline separates the middle zone and lower zone of DCE zone. Similarly, the median of uppermost co-ordinate of the components in the DCE zone generates a horizontal line. This horizontal line, called top line, separates the middle and upper zones of the DCE zone. The subscripts in a DCE zone belong to lower-half of the middle zone and lower zone whereas the superscripts belong to upper zone and upper-half of the middle zone. Based on the location of the components in a DCE zone we have detected the subscripts and superscripts.

D. Auto correction of reactants and products in chemical equations

E. Generation of ground truth data in PDF format

III. Experimental Result

Here goes the results

IV. Conclusion

The conclusion goes here.

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.