Extraction of NMR Spectra and Structural Data from Documents for Semantic Representation and Reuse

William Brouwer
Department of Chemistry
Pennsylvania State University
wjb19@psu.edu

Mark Borkum
School of Chemistry
University of Southampton
m.i.borkum@soton.ac.uk

Abstract

For at least fifty years, liquid state Nuclear Magnetic Resonance (NMR) has served as an important analytical technique in studying local atomic bonding information. Thus, a vast amount of data of interest to the chemist and crystallographer resides in archived documents, containing liquid state NMR spectra and accompanying molecular structures. These structures are determined on the basis of chemical shift information from spectra using well established empirical rules. The combined wealth of information represented visually in the spectra and molecules precludes straightforward inclusion in a traditional database. Given its value to the researcher, work by this group is being dedicated to automatic extraction of spectral and molecular information from documents, for conversion to a machine-readable format and incorporation into a database.

1 Introduction

Since its inception, Nuclear Magnetic Resonance[1] has become ubiquitous as a spectroscopic tool, particularly in organic chemistry. Commercial NMR spectrometers provide means for processing and analysing data, often developing 'peak lists' which may then be mapped to molecular structural information, using empirical rules and calculation. Typically, in publications, both molecules and spectra are represented using a two-dimensional vector format and published as a PDF document (Figure 1).

The benefits in extracting and storing this data lie chiefly in reuse for computer assisted structure determination [1], particularly with regards to complex methods [2]. There is to date a deluge of readily available spectroscopic data for this task, and this work aims to make this data available when it would otherwise be cast aside. To the best of our knowledge, this is a unique research contribution relying on both traditional and novel algorithms.

Document analysis and conversion are heavily explorer areas; recent efforts for instance have been devoted to conversion between different schema [3] for object reuse and exchange [4]. Much work has also been dedicated to information

[1] In this technique, nuclei with non-zero spin couple with a static magnetic field to form the Zeeman energy levels. Probe nuclei are stimulated at the resonant Larmor frequency between energy levels, the system response acquired and absorption frequency spectrum analysed. Common probes in the liquid state include ^{1}H and ^{13}C nuclei; the periodic table is filled with potential candidates, each subject to internal interactions which are perturbations to the Zeeman energy levels. Rapid tumbling motion in liquids eliminates anisotropic line broadening mechanisms, leaving only isotropic interactions which give rise to peak shifts from the Larmor frequency in the spectrum. Subsequently, the positions of lines in a liquid state NMR spectrum are highly sensitive to local bonding details, and further the integrated intensity is directly related to the proportion of spins in a particular chemical environment.

extraction from scientific documents, for example keyword extraction [5]. The work presented here is both concerned with information extraction and document conversion, from a fixed-layout, two-dimensional vector format to a machine-readable, semantically-rich form.

2 Preliminary

Vector documents in general consist of various drawing statements.[2] Objects within documents that pertain to spectra amplitude data (peaks) are composed of contiguous sections of lines and/or polylines, while the individual tics of the scale are vertical lines, and the axis a single horizontal line. The scale maps the chemical shift in units of parts per million (ppm) to peak positions in the spectra, by virtue of text labels anchored to vertical tics. Text characters are drawn using a combination of position, bounding box and line drawing commands. Recognition and extraction in this work is devoted to spectra amplitude data (hereafter simply referred to as spectra) and spectra scales. Figure 2 gives a representation of the overall process. As a first step towards semantic representation of the extracted data, spectra are modeled and peak list developed, and along with raw data, are stored in text files.

3 Method

Initial steps identify the various features of interest; lines, polylines and text character statements, along with their record numbers in the source file. Subsequent steps identify sets of ‘connected’ features by detecting intersecting bounding boxes. Heuristics are applied to classify sets as objects. Peak shifts and/or integrated intensities may also be present in the form of lines, polylines and text. In either case, this information is generally incomplete, and rather than extract it alongside spectra, the latter is modeled using a Gaussian mixture model and the full peak list generated automatically [6].

4 Conclusions and Future Work

There is an abundance of readily available chemistry data, in the form of NMR spectra and accompanying molecules, existing in PDF documents. This data is invaluable and of high quality, yet is essentially unusable for further modeling and analysis. The work describe herein is dedicated to the complete extraction of this data. At this stage, text character, line and polyline features are identified and removed from PDF source files and assembled into the objects of interest using novel and traditional algorithms. Peak lists are generated automatically from appropriately scaled spectra, making them highly amenable to reuse. Further work is being dedicated to the semantic representation of the extracted molecule and spectral data [7], as well as incorporation of the extracted objects into the oreChem framework [4].

[2] For the purposes of this work: ‘Line’ refers to a 4-tuple of numbers $(x_1, y_1, \Delta x_1, \Delta y_1)$, with additional drawing commands such as ‘stroke’, ‘scale’ etc. ‘Polyline’ refers to multiple line segments and drawing statements, ie., $2n$-tuples where $n > 2$. In either case, initial x, y floats give the starting coordinates for line segments, and subsequent floats $\Delta x_j, \Delta y_j$ correspond to displacements from the last position.

References

[1] M. Elyashberg, K. Blinov, S. Molodtsov, Y. Smurnyy, A.J. Williams, and T. Churanova. Computer-assisted methods for molecular structure elucidation: realizing a spectroscopist's dream. *Journal of Cheminformatics*, 1(1):3, 2009.

[2] GM Lipkind, AS Shashkov, YA Knirel, EV Vinogradov, and NK Kochetkov. A computer-assisted structural analysis of regular polysaccharides on the basis of 13c-nmr data. *Carbohydrate research*, 175(1):59, 1988.

[3] A. Boukottaya and C. Vanoirbeek. Schema matching for transforming structured documents. In *Proceedings of the 2005 ACM symposium on Document engineering*, pages 101–110. ACM New York, NY, USA, 2005.

[4] C. Lagoze. The orechem project: Integrating chemistry scholarship with the semantic web. *WebSci'09*, 2009.

[5] N. Kumar and K. Srinathan. Automatic keyphrase extraction from scientific documents using n-gram filtration technique. *DocEng'08*, 2008.

[6] H.N.B. Moseley, N. Riaz, J.M. Aramini, T. Szyperski, and G.T. Montelione. A generalized approach to automated nmr peak list editing: application to reduced dimensionality triple resonance spectra. *Journal of Magnetic Resonance*, 170(2):263–277, 2004.

[7] P. Murray-Rust, H.S. Rzepa, and B.J. Whitaker. The world-wide web as a chemical information tool. *Chemical Society Reviews*, 26(1):1–10, 1997.

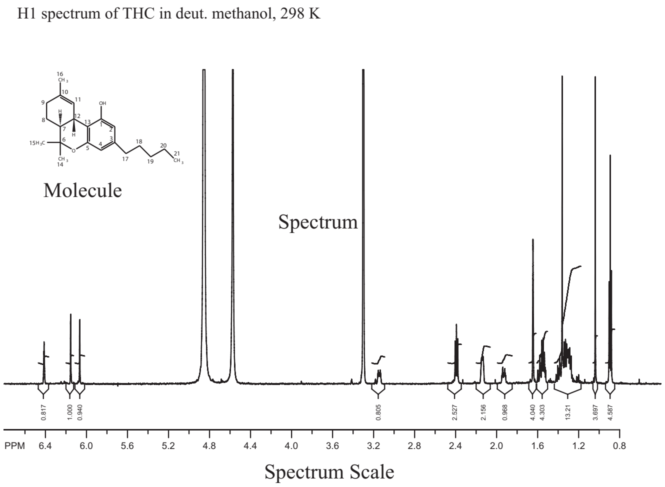

Figure 1. Image of high resolution NMR spectrum with accompanying molecule

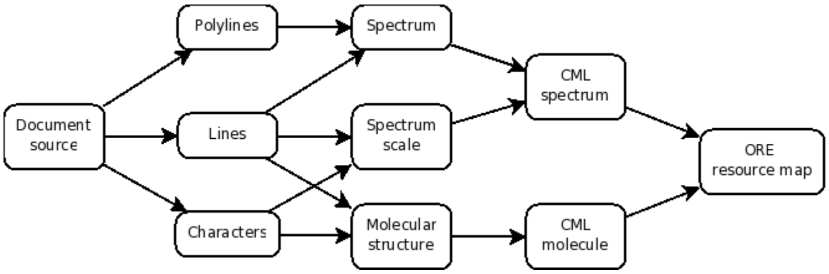

Figure 2. Extraction and enrichment of objects of interest from vector documents