学術論文からのポリマー溶解性データの自動抽出

岡 博之[1], 吉澤 篤志[1], 進藤 裕之[2,3], 松本 裕治[2,3], 石井 真史[1]
物材機構(NIMS)[1], 奈良先端大(NAIST)[2], 理研 AIP[3]

{OKA.Hiroyuki, YOSHIZAWA.Atushi, ISHII.Masashi}@nims.go.jp
{matsu, shindo}@is.naist.jp

1. はじめに

NIMS では以前からポリマーデータを学術論文から収集しており、それらをデータベースで管理・公開し[1]、また Materials Informatics に活用している。データ抽出は人手で行っているが、最近の出版論文数の増加、テキストマイニング技術の向上などから、コンピュータを用いた自動抽出も試みている。対象とするポリマーデータの種類はおよそ 100 あるが、これらの多くは数値データであることから、論文中の図や表にまとめられていることが多い。しかし、ポリマーの溶解性については、数値でなく、溶媒名とそれへの溶解性(例えば、"soluble"、"insoluble"など)で記述されているため、本文に記載されていることが多い。そこで本研究では、本文からポリマーの溶解性データを自動抽出するために、ポリマー名とその良溶媒名を関係づける関係抽出を行うことを検討した。溶解性データとしては、良溶媒だけでなく、貧溶媒についても抽出する必要があるが、本研究では、問題を簡単にするためにも、まずは良溶媒についてのみ行った。ポリマーがある溶媒に可溶な場合、それを表す文章にはポリマー名、溶媒名とともに"dissolve"や"soluble"などの単語が記載されていることが通常である。また、ポリマーを溶媒に溶解させて行う作業(粘度測定やフィルム作成など)の文章では、そこに記載の溶媒名は良溶媒名とすることができる。本研究では、ポリマー論文で見られるこのような傾向を考慮して、関係抽出をルールベースで行うことを検討した。

一方、ポリマーの溶解性は図1のように表にまとめられている場合もある。表からの自動データ抽出は、論文の xml ファイルを使用すれば可能である。我々は以前に、表からポリマーデータを抽出する方法を検討しており、それについての報告も行っている[2]。そこでは、表の行列番地を利用することで、例えば図1でいうと、"PI1"-"NMP"-"++"(それぞれ、ポリマー名－溶媒名－溶解性を示す)のような3つの固有表現がセットになったデータを抽出できる方法を示した。表中のポリマー名と物性名の予測が必要であったが、それぞれ深層学習とキーワードマッチングによって行えるようにし、結果、多くの論文から、多くのポリマーデータを精度良く抽出できることを示した。本研究では、その技術を用いて、表からの溶解性データの自動抽出も行った。表では、図1で見られるように、溶解性が"+"や"-"などのマーカーで示されていることが多い。そのため、この溶解性マーカーを利用することで抽出を行った。本文中のポリマー名－良溶媒名の関係抽出とともに、検討を行った結果をここに報告する。

TABLE 2 Solubility Data of the Polyimides Synthesized in this Work

PI	Solvents[a]								
	NMP	DMAc	DMF	DMSO	Toluene	Xylene	$CHCl_3$	THF	Acetone
PI1	++	++	++	++	±	+	++	++	+
PI2	++	++	++	++	++	++	++	++	±
PI3	++	++	++	++	±	++	++	++	+
PI4	++	++	+	+	++	++	++	++	±

[a] Solubility: (++) soluble at room temperature; (+) soluble upon heating at 60 °C; (±) partially soluble or swells.

図 1 表中のポリマー溶解性データの記載例(J. Polym. Sci. A: Polym. Chem., 2015, 53, 479-488.から抜粋).

2. 実験

プログラム作成はすべてPython 3を用いて行った。

2-1. 本文からのポリマー名－良溶媒名の関係抽出

2-1-1. 使用論文

論文は英語論文である Macromolecules(出版社:American Chemical Society, 2016 年分から 63 論文)を使用した。ファイルは xml 形式のものを使用した。

2-1-2. 関係抽出の評価用データの作成

関係抽出を行うに当たって、その精度を調べる必要があるために、まずは評価用データを作成した。文章中のポリマー名と溶媒名にアノテーションを行い、さらに、可溶性でこれらの間を関係づける

箇所にもアノテーションを行った。作業の最初は、人手労力を省くためにも、プログラムを用いて大まかに行った。このとき、ポリマー名の多くは”poly”から始まることなどを考慮して、ルールベースのプログラムを作成して行った。これについては、以前の表からのポリマーデータ抽出で、その詳細を報告しているので[2]、ここでは省略する。溶媒名については、溶媒名辞書をあらかじめ作成しておき、辞書マッチングによって行った。溶媒名には表記ゆれがあることが多く、例えば、クロロホルムは英語論文で”chloroform”または”$CHCl_3$”のどちらかで表現されることが多い。また、クロロホルムの場合、重水素化溶媒である”chloroform-*d*”または”CD Cl_3”がポリマー論文中では頻出であるが、重水素化の有無で溶解性が変わることはないので、クロロホルムの表記ゆれとして登録するようにした。他の溶媒についても同様に行い、合計で135の溶媒を登録した。こうして作成した溶媒名辞書を用いて、マッチングによって溶媒名のアノテーションを自動で行った。この後、人手による修正を行い、このとき関係するポリマー名－溶媒名のアノテーションも行った。この人手作業ではHTMLAnno[3]を用いた。この作業後、アノテーションデータを図2のように変換した。

116	.	O	_	_
117	The	O	_	_
118	more	O	_	_
119	polar	O	_	_
120	Py-PC1A	S-polymer ★	_	_
121	was	O	_	_
122	soluble ★	O	_	_
123	in ★	O	_	_
124	THF	S-solvent ★	120-poly-sol	120-poly-sol TP
125	.	O	_	_
126	DMF	S-solvent ★	120-poly-sol	120-poly-sol TP
127	.	O	_	_
128	and	O	_	_
129	DMSO	S-solvent ★	120-poly-sol	120-poly-sol TP
130	but	O	_	_
131	not	O	_	_
132	in	O	_	_
133	toluene	S-solvent ★	_	120-poly-sol FP
134	.	O	_	_
135	All	O	_	_

図2　関係抽出用データ．左から1列目：トークン ID、2列目：トークン、3列目：ポリマー名および溶媒名の正解タグ、4列目：ポリマー名－良溶媒名間の関係タグ（正解）、5列目：本研究でのルールベース関係抽出によって予測された関係タグ．赤字：ポリマー名、淡青字：溶媒名、緑星印：ルールベース関係抽出で条件を満たしている箇所。TP（青）：True Positive、FP（橙）：False Positive.

この作業では、ポリマー名および溶媒名に BIOES タグを用いて正解タグを付け（左から3列目）、また、それぞれを区別するために、後ろに”-polymer”および”-solvent”を付けた。次に、ポリマー名と良溶媒名の関係については、溶媒名の所で、関係するポリマー名のトークン ID（1列目）とその後に”-poly-sol”を付けることで関係タグを付けた（4列目）。図2の場合、トークン ID が 124、126 および 129 の溶媒が 120 のポリマーの良溶媒となるので、これら 3 つの溶媒名の所に”120-poly-sol”と正解の関係タグを付けた。なお、133 の”toluene”は貧溶媒であるため、タグ付けは行っていない。こうして関係抽出用の評価データを作成したが、人手労力を省くためにも、論文のすべての文章を用いるのではなく、溶解性に関する内容が記載されている段落のみを用いた。合計で 199 段落分（ポリマー名－良溶媒名を関係づけた数は 392）の評価用データを作成した。

2-1-3. ルールベースによるポリマー名－良溶媒名の関係抽出

本研究の関係抽出では、ポリマー名および溶媒名の固有表現認識の検討は行わず、評価用データの正解タグを利用して、ポリマー名－良溶媒名間の関係を取ることを主に行った。これを行うルールとして、1文内でポリマー名および溶媒名を含み、かつ、可溶性に関連する単語（2つ）を含んでいる場合に、その1文内のポリマー名と溶媒名はすべて関係があるとした。なお、1文の範囲はピリオドからピリオドまでとした。また、可溶性に関連する単語（2つ）は、表1の Index term 1 を用いて、部分マッチングされる単語と、Index term 2 の”in”、”from”、”using”の3つのうちどれか1つとした。Index term 1 での部分マッチングでは、”dissolve”や”soluble”に加え、”spin-coat”や”viscosity”などポリマー溶解を前提とする作業を表す語もマッチングできるようになっている。”in”、”from”、”using”はこれらの語と共起していることが多いため、ルールとして加えた。

表1　溶解性関連文章の検索用 Index term

Index term 1	solution, solvent, dissol, solubility, soluble, dilute, film, spin-coat, cast, concentrat, NMR, spectr, GPC, viscosity, measur, blend, prepar, precipitat
Index term 2	in, from, using

このようなルールによって関係抽出を行い、その結果を図2の5列目に、4列目と同様に、関係を表すタグで与えるようにした。図2では、116 と 134 のピリオドから、117～134 のトークンが1文とみな

され、その中で、ポリマー名と溶媒名が、3列目の正解タグから、120 と 124、126、129 および 133 のトークンに特定される。そして、可溶性に関連する単語は122の”soluble”と123の”in”が表1のIndex term 1 および 2 によってマッチングされ、結果、117～134 の1文内のポリマー名と溶媒名はすべて関係があるということになる。5列目の関係タグと4列目の正解の関係タグとを比較することで、124、126 および 129 の所は True Positive（TP）、133 の所は False Positive（FP）と判定させた。このようにして、ルールベースによる関係抽出とその評価を行った。

2-2. 表からのポリマー溶解性データの抽出

2-2-1. 使用論文

論文は英語論文である Journal of Polymer Science Part A: Polymer Chemistry（出版社：Wiley, 2015-2016 年分から 100 論文）を使用した。ファイルは xml 形式のものを使用した。

2-2-2. 抽出用データ作成および抽出方法

表からのデータ抽出を行うために、表の行列番地を利用して、図 1 での”PI1”－”NMP”－”++”のような3つの固有表現のセットを抜き出し、それを、図3の左から1列目のように、縦一列に並べることを行った。この作業については、以前の発表で報告しているため[2]、ここでは詳細を省略する。

PI1	S	O	O
NMP	O	S	O
++	O	O	S
PI1	S	O	O
DMAc	O	S	O
++	O	O	S
PI1	S	O	O
DMF	O	S	O
++	O	O	S
PI1	S	O	O
DMSO	O	S	O
++	O	O	S
	ポリマー名	溶媒名	溶解性マーカー

図3 ポリマー名、溶媒名および溶解性マーカーの予測結果. 左から1列目：表中のトークン、2列目：深層学習によるポリマー名の予測タグ、3列目：マッチングによる溶媒名の予測タグ、4列目：マッチングによる溶解性マーカーの予測タグ

次に、1列目のデータに対して、ポリマー名の予測を行った。これは深層学習を利用して行った。これについても以前に報告を行っているので[2]、詳細は省略する。結果は”S”または”O”のタグで与えるようにしており（図3の2列目）、それぞれポリマー名とそうでない場合を示す。次に、溶媒名の予測を行った。これは関係抽出のところで作成した溶媒名辞書とのマッチングによって行った。これも結果のタグを”S”または”O”で与えるようにし（図3の3列目）、それぞれ溶媒名とそうでない場合を示す。最後に、溶解性を表すものかどうかの予測を行った。溶解性データの表では、溶解性を表す方法として、図1の表で見られるように、”+”や”−”、また、”S”や”I”（それぞれ soluble、insoluble を意味する）などのマーカーで表現されていることが多い。そこで、溶解性マーカーをあらかじめ調べてそのリストを作っておき、マッチングによって、溶解性の予測を行った。具体的には表2にリストしたマーカーを用いて行った。プラス、マイナスの記号では、これらが2つ連続して記載されている場合が多くあるが、このとき、その間にスペースがないとき（”++”）とあるとき（”+ +”）があり、その揺らぎも溶解性マーカーとして登録した。

表2 本研究で用いた溶解性マーカー

+, −, ±, +++, ++, +−, −+, −−, + +, + −, − +, − −, S, I, +h, Soluble, Insoluble, Swelling

PI1	NMP	++
PI1	DMAc	++
PI1	DMF	++
PI1	DMSO	++
PI1	Toluene	±
PI1	Xylene	+
PI1	CHC13	++
PI1	THF	++
PI1	Acetone	+
PI2	NMP	++
・	・	・
・	・	・

図4 表から抽出した溶解性データ（最終出力）. 1列目：（表記載）ポリマー名、2列目：（表記載）溶媒名、3列目：溶解性マーカー.

表2の溶解性マーカーを用いて、マッチングを行った結果を”S”または”O”のタグで与えるようにした（図3の4列目）。それぞれ溶解性マーカーとそうでない場合を示す。このようにして3つの固有表現の予測を行った後、結果のタグを縦に並べたデータを3行毎にチェックし、溶解性データとして抽出できるかの判定を行った。例えば、図3の赤枠内で、”PI1”、”DMAc”および”++”の”S”のタグがそれぞれポリマー名、溶媒名および溶解性マーカーにのみ付いているので、この赤枠内の3行は、ポリマー名－溶媒名－溶解性の3つの固有表現がセットになったものと判定できる。この場合は溶解性データとみなすことができるので、このセットを抽出し、最終的には図4の形で出力するようにした。なお、この抽出作業を行う前に、溶解性マ

ーカーの”–”(マイナス)は、ハイフンとして(データなしという意味で)使われていることも多いので、抽出時の False Positive(FP)を避けるために、数ある表の中から溶解性データの表をプレスクリーニングする作業も行った。溶解性データの表では、タイトルに”solubility”あるいは”solubilities”の単語が含まれていることが多いので、これらとマッチングすること、表中に溶媒名が1つ以上記載されていること、さらに、溶解性マーカーが表中のデータ数に対して 1 割以上含まれていることを条件として、作業を行った。この表プレスクリーニング精度と溶解性データの抽出精度の評価は人手によって行った。

3. 結果と考察

3-1. 本文からの溶解性データの抽出

評価用データを用いて、ルールベースで関係抽出を行った結果、True Positive(TP)、False Positive(FP)および False Negative(FN)の数はそれぞれ 358、263 および 34 であった。これらの値から、Precision、Recall および F 値がそれぞれ 0.577、0.913 および 0.707 と算出された。FP が多く、FN が少ない結果であったが、FP が多くなった原因の一つとして、ピリオドでの文区切りが良くないと考えられた。文途中で”and”や”but”などの接続詞が入ると、それ以降、溶解性に関する文章でなくなっている場合があり、このようなとき、その部分に記載の溶媒名は誤って抽出されていた。これらについては改善を進めている。

3-2. 表からの溶解性データの抽出

用いた 100 論文中には、260 個の表があり、その中で溶解性に関する表は 12 個であった。表プレスクリーニングでは、これら 12 個を全て自動抽出できていた。この中から抽出したポリマー名－溶媒名－溶解性データで、TP、FP および FN をチェックしたところ、数はそれぞれ 351、0 および 151 であった。これらの値から、Precision、Recall および F 値はそれぞれ 1、0.699 および、0.850 と算出された。FP の数が 0 であった理由は、全ての表がプレスクリーニングできていたことと、今回用いた論文中には、粘度などの他のデータが一緒に記載されているような表はなかったことに依った。FN については、ポリマー名の抽出漏れが多くあった。その中には、ポリマー名が”8a”などのように、著者によって定義されたサンプル ID で表現されているものが多かった。溶解性データの表 12 個中 3 個で、ポリマー名がすべてサンプル ID で表現されており、そのため、これらのデータ(80 個)はすべて FN となってしまった。このような表現のときの機械によるポリマー名予測はまだ困難であり、これからの課題の一つである。また、辞書に未登録の溶媒名によって FN となったケースもあり、辞書の編纂の高品質化は課題の一つである。この編纂では”DGTE”(diethylene glycol dimethyl ether)といった溶媒の慣用名の登録が特に重要である。最後に、溶解性マーカーで FN となった例では、”U”や”U*”の表現のものがあり、これらは UCST タイプの溶解性(ある温度以上で可溶となる)を表しているが、このようなマーカーは登録していなかった。

4. おわりに

学術論文からポリマー溶解性データの自動抽出を本文と表からの2通りで行った。本文の方ではポリマー名－良溶媒名の関係抽出をルールベースで行い、その精度(F 値)は約 0.71 であった。表からの抽出では、以前に報告した方法を利用して行ったところ、抽出精度(F 値)は 0.85 であった。両方とも、ルールの改善や辞書登録数の増加などで、精度の改善が期待できる。発表当日は改善後の結果を含めて報告を行う予定である。

参考文献および URL

[1] PoLyInfo, https://polymer.nims.go.jp/

[2] H.Oka et al., “Automatic extraction of polymer data from tables in xml”, Third International Workshop on SCIentific DOCument Analysis (SCIDOCA2018), 慶応義塾大学 日吉キャンパス(横浜市), 2018 年 11 月 12–13 日.

[3] NAIST 開発アノーテションツール, https://github.com/paperai/htmlanno