千一問の質問における型

亀田 尭宙

ディスカッション・ペーパー『『カラム』の時代Ⅵ――近代マレー・ムスリムの日常生活２』では「カラムデータベースにおける理解支援――展望と周辺技術」と題し、Entity Linkingやレトリック構造の抽出について述べた。そのうち、レトリック構造の分析に手を付けるにあたって、今回は、まず『カラム』のQAコーナー「1001 Masalah（千一問）」から質問の型を見つけ出すことに取り組んだ。

現在、千一問のコーナーを日本語に訳す作業が進んでいるが、データがそろっているマレー語の方を分析の対象とした。一方で、以下の分析を構想したのは日本語訳を読んでいて着想を得た部分が大きい。質問の多くは、何かの行為がイスラム法的に合法か違法かを問うており（例：「ネクタイ、帽子や膝の見えるズボンを着用した場合、法的にはどうなりますか（Q.67）」[*Qalam* 1951.3:15]）、他には事実知識を問うものもあれば（例：「国連会議に参加する代表は各国何人いますか（Q.52）」[*Qalam* 1951.2:41]）、人生相談もある（例：「女性はいつ結婚するのが最も良いですか（Q.48）」[*Qalam* 1951.2:40]）。回答のレトリックを分析する際にも、文単位でどういう役割の文かを判別しなければならないが、比較的短くタイプがはっきりしていると考えられる質問文から型を機械的に判別することに、まず取り組むことにした。

本稿の流れは以下の通りである。

- 語の並びに着目するため、前処理として質問文を正規化した語の並びに変換する
- 各語の出現頻度など統計値を取り、それに基づいて、型の骨格を成す頻出語の列を得る
- 頻出語列の部分列に対し「型らしさ」を測る指標を作り、適用することで型を探す

前処理――各質問文を語幹の列としてデータ化する

元のデータはエクセルで表現されており、中には日本語も特殊文字も多く含まれている。これを文字化けさせずにプログラム処理用に取り出すため、xlsx2csv[1]を用いてcsvに変換した。

さらに、取り出したものの中から、「東南アジア逐次刊行物総合目録データベース」[2]の副産物として公開されている音標変換表を用いて、特殊文字を一般的なアルファベットに変換した。この中には“ā”を“a”に変換するといった規則が含まれており、多くの特殊文字を変換することができた。文字の下に記号がついている特殊文字については、元の変換表に含まれていなかったため、“ṣ“を“s”に変換するルールなどを追加し、対応した。

次に、アルファベットとハイフンやアポストロフィだけになった語の列に対し、語幹を取り出す作業（Stemming、これを行うツールをStemmerと呼ぶ）を行った。カラムで用いられているマレー語には、ke-、per-、-nyaといった接頭辞や接尾辞が頻繁に出現し、それらを取り除いた語幹に着目することで、似た意味の単語をまとめて扱うことができると考えたからである。具体的には、全文検索エンジンApache Luceneに実装されているプログラムを用いた。Luceneには、インドネシア語を分析するためのモジュールorg.apache.lucene.analysis.id[3]が含まれており、この中のStemmerは[Tala 2003][4]に基づいて実装されている。本来はマレー語のStemmerを利用すべきだと考えたが、この両言語についてStemmingのレベルではほぼ違いが無く、一方でマレー語のStemmingに関する詳細を記述した論文やツールを見つけられな

1）https://github.com/cm3/xlsx2csv Dilshod Temirkhodjaev氏の xlsx2csv という変換ツールの文字コードの扱いについて私が修正を施したもの。
2）http://www.cseas.kyoto-u.ac.jp/info/db/sealib/
3）http://lucene.apache.org/core/5_4_1/analyzers-common/org/apache/lucene/analysis/id/package-summary.html
4）Fadillah Z Tala: “A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia”, 2003 http://www.illc.uva.nl/Research/Publications/Reports/MoL-2003-02.text.pdf

かったため、代用することにした。

ここまでで、各質問文は、語幹の列として表現することができた。

例[*Qalam* 1954.8:5][5)]

Meminta sedikit penjelasan tentang binatang sembelihan – qurbān yang biasa dikerjakan oleh orang Islam pada Hari Raya Haji.

→ ["inta", "sedikit", "jelas", "tentang", "binatang", "sembelih", "qurban", "yang", "biasa", "kerja", "oleh", "orang", "islam", "pada", "hari", "raya", "haji"]

頻出語列を得る

全質問文に対して、語幹の列を集計したところ、表1のような結果を得た。上位の多くを機能語(人称代名詞や前置詞など語彙的な意味を持たず統語的な機能のみを持つ語)が占める中、hukum, islam, agamaといった内容語も見られ、その多くはイスラム教に関連している。今回は40回以上出現している42語を頻出語とみなし、各質問文の列から相当する部分列を抜き出した。

「型らしさ」の測定

例えば、["apa", "hukum"]という頻出語列は非常に「型らしい」。それ自体が60回出現するだけでなく、冒頭で述べたような人間が読んで感じる質問の型と対応するからだ(例Q36:"Apa hukum taklik"[*Qalam* 1951.1:33]、日本語:「タッリーク(taklik:婚姻の際に交わされた離婚条件などの契約)は、法的にはどうなりますか」)。

一方で出現頻度がそれなりに多くても型とは言いづらいものもある。例えば、["yang", "saya"]という頻出語列は12回出現するが、4回しか出現しない["bagaimana", "hukum", "orang", "yang"]という語の列に比べて、質問文を読んだときに共通点を発見しづらい。

この「型らしさ」を以下のように測定した。

文の長さは出現する語の選択に影響を与えないと仮定し、各語を目としたサイコロが複数回振られる(独立性の仮定)ことで文の中の語が決まるというモデルを採用する。これはLatent Dirichlet Allocation[6)]のような一般的なトピックモデルと共通した仮定になっている。

表1 語幹の出現数 上位20

語幹	出現数
yang	452
itu	302
dan	269
di	250
ada	223
dengan	195
dalam	170
tidak	144
atau	138
hukum	135
islam	130
saya	115
orang	111
seorang	106
apa	106
pada	93
agama	86
sembahyang	85
oleh	82
daripada	82

それぞれの頻出語w_i(i=1, 2, 3,…)、および頻出語でない語w'について出現確率$p(w_i)$や$p(w')$が存在するとする。これらの確率の総和は1である。この多項分布上で文中の語数に応じた回数、独立に語が選択され、頻出語の列$[f_1, f_2, f_3,\ldots]$が得られる。n語で構成される文の列内に特定の頻出語部分列、たとえば$[w_1, w_2]$が存在する可能性$q([w_1, w_2], n)$は次の式で求まる。

$$q([w_1, w_2], n) = p(w_1)\cdot(1-(1-p(w_2))^{(n-1)}) + (1-p(w_1))\cdot p(w_1)\cdot(1-(1-p(w_2))^{(n-2)}) + (1-p(w_1))^2\cdot p(w_1)\cdot(1-(1-p(w_2))^{(n-3)}) \ldots + (1-p(w_1))^{(n-2)}\cdot p(w_1)\cdot p(w_2)$$

$[w_1, w_2, w_3]$ならば、

$$q([w_1, w_2, w_3], n) = p(w_1)\cdot q([w_2, w_3], n-1) + (1-p(w_1))\cdot p(w_1)\cdot q([w_2, w_3], n-2) + (1-p(w_1))^2\cdot p(w_1)\cdot q([w_2, w_3], n-3) \ldots + (1-p(w_1))^{(n-3)}\cdot p(w_1)\cdot q([w_2, w_3], 2)$$

ちなみに、$q([w_1], n)$は

$$q([w_1], n) = p(w_1) + (1-p(w_1))\cdot p(w_1) + (1-p(w_1))^2\cdot p(w_1) \ldots + (1-p(w_1))^{n-1}\cdot p(w_1)$$

5) この例では、冒頭のMemintaのStemmingに誤りがあり、本来はmintaとするべきであるが、今回はStemmerの改善を行わなかったため、そのままの結果を示している。

6) Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3, 2003, pp.993-1022. で提案されている手法。

$= p(w_1) \cdot (1-(1-p(w_1))^n) / (1-(1-p(w_1)))$
$= 1-(1-p(w_1))^n$
と簡素になる。

［w_1, w_2］と［w_1, w_2, w_3］の関係で示したように、この式は再帰的アルゴリズムとして実装することができる。このqで算出された確率は文に型などなく独立に語が出現するという仮定の下の確率であるので、実際の出現数がこの確率よりどれだけ多いかということを測定することが「型らしさ」の測定となる。そこで、文に当該の頻出語列が出現したか／しないかについての二項分布の下側累積確率を全文について求めた。

その結果、2語だと［“apa”, “hukum”］,［“agama”, “islam”］といった頻出語列がほぼ100%[7]型として共起しているという結果が得られ、4語の［“bagaimana”, “hukum”, “orang”, “yang”］(4件)もほぼ100%という結果になった。一方で、［“yang”, “saya”］(12件)は13.9%、［“hukum”, “ada”］(11件)は64.1%となり、［“bagaimana”, “hukum”］(12件)がほぼ100%になったことを考えても、出現件数に比して型らしさは低いと考えられる。

まとめと今後の展望

千一問の中から、質問に共通して現れる頻出語の列を見つけ出し、その型らしさを測定した。いくつかの例から、当初想定していた型を抽出することはできたが、全体の評価はまだ行っておらず、実際の例と比較しながら手法の妥当性を検討する必要がある。また、本来は、モデルの時点で共起の確率を考慮する必要がある。今回は質問文のみをデータとして用いたため、学習データの量の観点からそのような複雑なモデルを採用しなかったが、Wikipediaなどの外部データを適切に使うことで、複雑なモデルを学習することも可能だと考えられる。

既に課題として見つかっているのが、［“agama”, “islam”］のように、確かにこの質問文に特徴的な共起であるが質問の型とは言えないペアも、型として高く評価されてしまう点があり、少なくとも1つ以上の機能語を含むことを条件とすることを検討している。

7）プログラムの精度の問題で、小数点以下10桁までででは100%とみなされた。