E-037

社内文書からビジネスドメインの知識グラフを構築する手法の検討
Domain-specific knowledge graphs based on in-house business documents

ホアン　レアン† 齋藤 靖二† 佐野 恵一† 鈴木 健太郎†
Hoang Le Anh Saito Seiji Sano Keiichi Suzuki Kentaro

1. はじめに

近年、DX(デジタルトランスフォーメーション)の発展は著しい。企業では営業、技術、製造から販売・保守にいたるまでデジタルでつながった業務に変革され、効率化が進んでいる。いくつかある DX ソリューションの中で、ノウハウ・知識が記載された社内文書を登録し、検索によって欲しい文書に早くたどり着ける検索プラットフォームは業務効率化とって非常に有用である[1]。しかしながら、適切な検索ワードを使用しないと所望の検索結果を得られず、繰り返し検索を行ってしまい時間を要してしまう。本課題の解決手段の 1 つとして、知識グラフを活用して適切な検索ワードを導出し、検索者の入力を支援することが考えられる。本稿では、このような検索プラットフォームの高機能化を目的に、社内文書からビジネスドメインの知識グラフを自動で構築する手法について検討を行った。

2. 知識グラフの活用シナリオ

図 1 に知識グラフを自動で構築し、検索プラットフォームの検索ワードを補完することで検索を支援するシナリオを示す。

検索プラットフォームには、営業・技術・品質・広報などビジネスドメインの文書が登録され、知識グラフは本文書をもとに自動構築する。構築した知識グラフを活用した場合、検索ワードを起点に知識グラフを辿ることで関連ワードを導出し検索ワードを追加する。これにより検索範囲を的確に絞り込め、欲しい文書に早く到達することができる。図 1 では品質対策を検索している例を示す。検索者が「破損」という検索ワードしか思いつかない場合、知識グラフから「圧力」や「漏洩」といったワードを補完することができる。1 つの時に比べ、3 つの関連したワードで検索することで、所望の検索結果を得やすくなる。

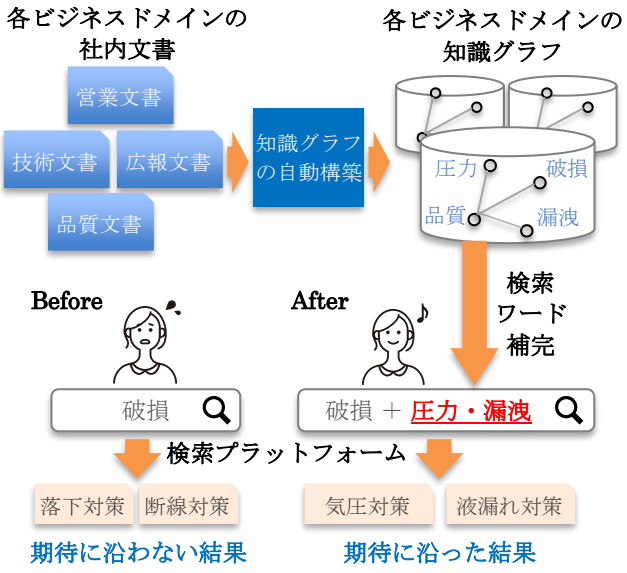

図 1 知識グラフの活用シナリオ

3. 知識グラフの構築

3.1 知識グラフの構造定義

第 2 章で述べた活用シナリオを実現するため、社内文書から構築する知識グラフを図 2 のような構造とした。社内文書に含まれる特徴語をノードとし、文書から抽出された特徴語同士の関係性をエッジとした（図 2）。

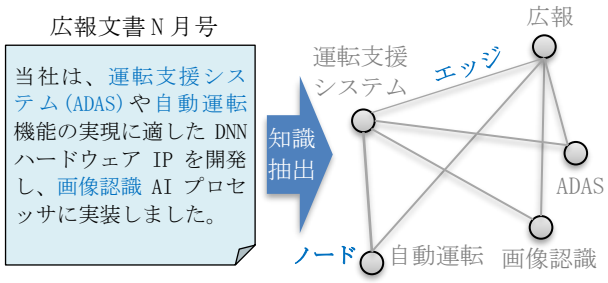

図 2 知識グラフの構造定義

知識グラフのノードとなる文書の特徴語を抽出する手順を図 3 に示す。最初に入力した文書のテキストデータを形態素解析によって単名詞に分解し、連続する単名詞を複合語として抽出する。次に中川らが提唱する FLR 法[2]を採用して対象となる複合語の重要度を算出し、その値が高い上位の複合語を特徴語として抽出した。

FLR 法において、L 単名詞から構成された複合語 CN=「$N_1N_2N_3\cdots N_L$」の重要度(FLR 値)は、(1)式で表される。

$$FLR(CN) = f(CN) \times LR(CN) \quad (1)$$

ここで $f(CN)$は文書内で複合語 CN が単独で出現する回数である。$LR(CN)$は重要度計算関数であり、(2)式に示す。

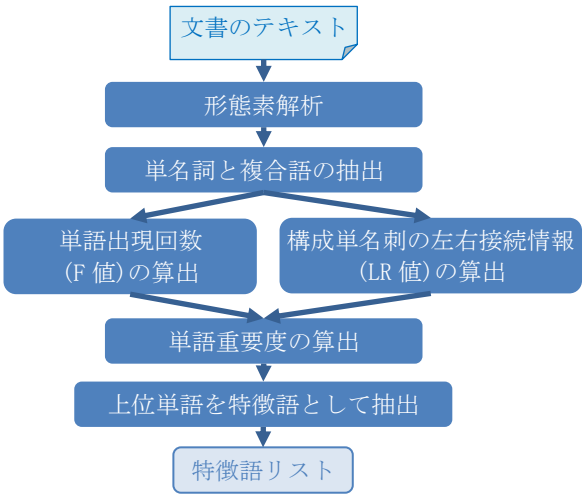

図 3 知識グラフのノードの抽出

† (株)東芝　Toshiba Corporation

複合語を構成する単名詞 N_i に対し、左側に単名詞が連接する頻度を $FL(N_i)$、右側に単名詞が連接する頻度を $FR(N_i)$ とする。単名詞 N_i に左右連接する頻度の相乗平均を求め、これを $LR(CN)$ する。このように相乗平均を取ることで、重要度が複合語 CN の長さ(L)に依存しない値となりうる。複合語が単独で多く出現し、かつ分解した単名詞も他の単名詞と多く連接するほど重要度が高くなる。この重要度が上位となる複合語を文書の特徴語とした。

$$LR(CN) = \left(\prod_{i=1}^{L}(FL(N_i)+1)\times(FR(N_i)+1)\right)^{\frac{1}{2L}} \quad (2)$$

3.2 知識グラフのエッジ抽出

特徴語となった単語同士の関係性を 2 つに分類し、それぞれエッジをとして採用した。

3.3.1 単語間共起関係からの単語間関係の抽出

文書にける、文章の節や段落を表すパラグラフやセンテンスにおける単語間の共起関係をもとに関係性を整理した。文書のテキストデータと抽出された特徴語を入力とする。同じ文書のセンテンスやパラグラフに含まれることから特徴語間共起関係を抽出する。次に対象のセンテンスやパラグラフの構文解析結果に基づいて抽出した単語間関係性を IS-A、PART-OF、INCLUDE、RELATED-TO に分類し、知識グラフのエッジとして表現する(図 4)。

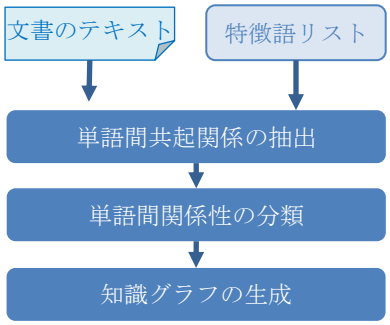

図 4 単語間共起関係からの単語間関係の抽出

3.3.2 文書全体の構成からの単語間関係の抽出

パラグラフやセンテンス単位だけでは、文書全体の構成や論旨をエッジに反映できないため、図 5 のような単語間関係の整理方法を立案した。最初に文書のテキストデータと文章の主題、論旨、単語の属性や接続関係等を表現する手掛かり語を入力として、例えば、課題と対策を示すような単語を抽出する。課題と対策に相当する単語同士を紐付けし、SOLVED-BY の単語間関係を有するとして整理する。整理した単語間関係は知識グラフのエッジとして扱う。

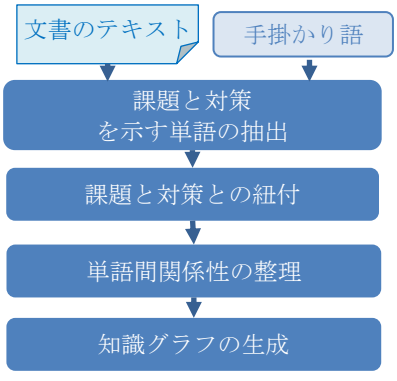

図 5 技術文章構成からの単語間関係の抽出

4. 検証

4.1 知識グラフ構築の試行結果

適用確認では、社内の 5 つのビジネスドメインに対して知識グラフ構築の試行を実施した。過去 5 年分の広報文書と技術報告書から合計約 25,000 ノードと 250,000 エッジから構成される知識グラフを生成した(表 1)。

表 1 知識グラフ試行の結果

ビジネスドメイン	文書件数	ノード数	エッジ数
A ドメイン	131	5275	17671
B ドメイン	350	6488	73006
C ドメイン	242	4160	40561
D ドメイン	190	4742	60139
E ドメイン	210	5016	58876

4.2 評価検証

全ノード数に対する有効ノード数の比率と、全エッジ数に対する有効エッジ数の比率で表す抽出精度を用いて、構築した知識グラフを評価した。ここで抽出された特徴語リストから文書と関係性の低い単語や、特徴語として扱われない単語を無効ノードとし、無効ノードに繋がったエッジを無効エッジとした。また、全ノードから無効ノードを除いたものを有効ノード数、同様に全エッジから無効エッジを除いたものを有効エッジ数とした。評価結果を表 2 に示す。ビジネスドメインによらず、ノードとエッジとも高い精度で抽出していることを確認できた。以上から、本手法を用いることで社内文書ごとに特徴語で構成された知識グラフを高い精度で、かつ自動で構築できると考える。

表 2 検証結果

ビジネスドメイン	無効ノード数	ノード抽出精度	エッジ抽出精度
A ドメイン	158	97%	98%
B ドメイン	130	98%	100%
C ドメイン	125	97%	99%
D ドメイン	190	96%	99%
E ドメイン	251	95%	99%

5. おわりに

社内文書からビジネスドメインの知識グラフを構築する手法を立案した。ビジネスドメインによらずノードとエッジとも高い精度で抽出していることが確認できた。今後、社内の情報検索システムの開発に活用し、検索機能の高度化を実現していく。

参考文献

[1] 佐野恵一,“階層的データ管理と複数データ領域の高速横断検索を実現する社内ナレッジシステム”, FIT2022 O-021
[2] 中川 裕志 et. al.,“出現頻度と連接頻度に基づく専門用語抽出”, 自然言語処理 2003 年 10 巻 1 号 p.27-45, 2003