Vol.2016-ICS-183 No.7
2016/3/16

日本語における依存構文と EMD に基づいた文間の関連度計算手法

賈珍磊[†1]　吉村枝里子[†2]　土屋誠司[†2]　渡部広一[†2]

概要：現代ではインターネットが普及しており，個人が簡単に電子化情報を収集できるようになっている．しかし，既存の検索システムは基本的には検索語の表記のみを活用するため，意味的には同じ内容の検索語でもユーザが入力する語によって検索結果が異なってしまう．そのため，利用者の要求に合った情報を探し出す必要性が高まっている．我々は単語から意味が近い単語を連想し，単語同士の関連や関係性を判断することができる．さらに，文章においては，我々は文章間の関連を連想することができる．文章の類似性を判断する機能を有したシステムは，辞書検索やウェブ検索など自然言語を処理するシステムへの応用が考えられる．手法の一つとして「概念ベースと Earth Mover’s Distance を用いた文書検索」という既存研究である．本研究では，文章に含まれる単語同士の関係性や意味だけではなく，文法規則も考慮する．そこで，依存構文と EMD 理論を用いた文間関連度計算手法を提案する．NTCIR という評価セットを用いて評価を行い，既存研究の「概念ベースと Earth Mover’s Distance を用いた文書検索」より精度が 10.6%が上がった．

キーワード：概念ベース，EMD，依存構文，類似度計算

Degree of Association between Japanese News by Dependency Structure and EMD

ZHENLEI JIA[†1]　ERIKO YOSHIMURA[†2]
SEIJI TSUCHIYA[†2]　HIROKAZU WATABE[†2]

1. はじめに

現代ではインターネットが普及しており，個人が簡単に電子化情報を収集できるようになっている．例えば，google[1]など検索システムの利用が一般的になっている．しかし，既存の検索システムは基本的には検索語の表記一致のみを活用するため，ユーザが望む検索結果が得られないことがある．そのため，利用者の要求に合った情報を探し出す必要性が高まっている．そこで，要求文と検索対象文の文間の類似性を定量化し,文を比較することによって，利用者の要求に近い情報を見つけ出すことが期待されている．

ユーザのニーズに合う検索結果の精度を高める方法として，単語から意味が近い単語を連想し，単語同士の関連や関係性を判断する手法と考えられる．文章の類似性を判断する機能を有したシステムは，辞書検索やウェブ検索など自然言語を処理するシステムへの応用が考えられる．そこで，文章間の類似性を定量化し，文章を比較することができるシステムが必要である．

文章においては表記が一致しない単語間においても互いに意味的な関連性を持っている．そこで単語間の関連性を表現するために構築された概念ベース（後述）を用いることで各単語の意味的な関連性を考慮した検索要求と検索文章間の類似性を判断する．

検索要求と検索文章間の類似性を判断する手法の 1 つとして Earth Mover’s Distance を用いた文間関連度計算方式[2]（以下，EMD）がある．概念ベースによって単語間の意味的な関連性を0.0から1.0までの数値として算出する．そして，単語間の意味的な関連性をもとに検索要求と検索対象との類似度を画像検索等の分野で注目されている EMD を用いることで単語間の意味的関連性を考慮して類似性を判断する．EMD とは距離尺度の 1 つであり，検索要求と検索対象 2 つの集合の距離を表すものである．

[2]で提案された過去に構築された EMD を用いた文間関連度計算方式では比較する二つの文章において，それぞれの単語は$tf \cdot idf$値[3]（後述）によって重要度をもつ.$tf \cdot idf$値は各単語の文章中における出現頻度に依存する．しかし，実際の文において，単語の重要度は出現頻度のみでは適切に表せないと考えられる．例えば,「私はきれいな嵐山に行った」という文では,「私」,「きれい」,「嵐山」,「行った」の単語が全部一回ずつ出現しているので，重みは同じとなる．しかし，実際には「私」,「きれい」よりも「嵐山」,「行った」という単語の方が重要であると考えら

†1　同志社大学大学院　理工学研究科
Doshisha University Graduate School of Science and Engineering
†2　同志社大学　理工学部
Doshisha University Department of Science and Engineering

れる．そこで，重要な単語に大きい重みをつけることで，より適切に文の意味を表せると思われる．

以上より，本研究では，文章に含まれる単語同士の関係性や意味だけではなく，文法規則も考慮する．そこで，日本語における依存構文[4]と EMD を用いた文間関連度計算手法を提案する．

2. 関連技術

2.1 概念ベース

概念ベース[5]とは，電子化された国語辞書や新聞文章などから機械的に構築された知識ベースである．概念ベースには約 9 万語の概念が収録されており，一つの概念に平均約 30 個の属性が存在する．例としてある概念Aはm個の属性a_iと重み$(w_i > 0)$の対によって次のように表現される．

$$\text{概念}A = \{(a_1,\ w_1),\ (a_2,\ w_2),\ \cdots,\ (a_m,\ w_m)\} \quad (1)$$

概念Aの意味定義を行う属性a_iを，概念Aの一次属性と呼ぶ．概念ベースの特徴として，属性を成す単語群も概念ベースの中で概念として定義されている点がある．つまり，属性a_iを概念とみなして更に属性を導くことができる．概念a_iから導かれた属性a_{ij}を，元の概念Aの二次属性と呼ぶ．概念ベースの例を表 1 にする．

表 1　概念ベースの例

概念	属性，重み
雪	(雪, 0.61) (雪掻き, 0.3) (粉雪, 0.27) …
雪掻き	(雪掻き, 0.16) (除雪, 0.14) (除雪, 0.14) …
粉雪	(粉雪, 0.23) (真っ白, 0.21) (氷点下, 0.2) …
⋮	⋮

概念からは任意の次元までの属性を一次，二次，三次，…, N次と導くことができる．このことより概念ベースは，N 次の属性による連鎖構造によって構成されていると言える．

2.2 一致度計算[5]

概念ベースを用いた単語間の関連性の定量化[6]は，基本的に語意の展開結果[6]を利用し数値として表す．何次属性まで展開するか，どの属性を用いるかによって値が変わってくるため，状況に応じてどのように計算するかが問題になってくる．本研究では，文間の類似度を求めるための単語間の関連性の定量化には一次属性までしか展開しない一致度計算を用いる．これは文書を概念と見立てた場合，索引語（後述）が一次属性となり，索引語の属性が二次属性となる．つまり，索引語の二次属性まで展開すると文書を概念とした場合の三次属性まで展開したこととなり雑音が増加し，概念（文書）とはかけ離れた語が計算に使用されてしまうためである．

索引語とは，文書内で「数」を除いた「名詞」，「形容詞」，「動詞」と定義する．

ある概念 A, Bにおいて，その属性をa_i，b_j，対応する重みをu_i，v_jとし，それぞれ属性が L 個，Q個$(L \leqq Q)$とすると，概念 A, Bはそれぞれ

$$A = \{(a_1,\ u_1),\ (a_2,\ u_2),\ \ldots,\ (a_L,\ u_L)\} \quad (2)$$

$$B = \{(b_1,\ v_1),\ (b_2,\ v_2),\ \ldots,\ (b_Q,\ v_Q)\} \quad (3)$$

となる．このとき，概念 A と概念 Bの属性一致度 $DoM(A, B)$を以下のように定義する．

$$DoM(A,B) = \sum_{a_i=b_j} \min(u_i, v_j) \quad (4)$$

$$\min(u_i, v_j) = \begin{cases} u_i (u_i \leq v_j) \\ v_j (u_i > v_j) \end{cases} \quad (5)$$

ここで，$a_i = b_j$は属性同士が一致した場合を示している．つまり，一致度とは概念 A と概念 Bそれぞれの属性の中で一致したものがあれば，同じ属性の重みにおいて小さい方の重みを選んで合わせた総和となる．これは，小さい方の重みは互いの属性の重みの共通部分となっているので，概念 A と概念 Bどちらにも有効な重みだと言える．ただし，各概念の重みの総和をそれぞれ 1 に正規化する．定義から明らかなように両概念の属性と重みの両方が完全に一致する場合には一致度は 1.0 となる．

2.3 $tf \cdot idf$

文章の文間関連度を求めるためには文章索引語に対する重み付けしなければならない．本稿では重み付けに$tf \cdot idf$を用いる．tfとは索引語頻度を意味し，索引語の網羅性を示す値である．文章中に索引語がどれだけ多く出現するかを示しており，何度も繰り返し使われる語は重要であると考えられる．文章d中に出現する索引語tの頻度を$tf(t,d)$と表す．また，文章d中に出現する総単語数sの頻度を$tf(t,d)$と表す．本稿では，$tf(t,d)$は文書dにおける索引語tの出現頻度である．ただし，$tf(t,d)$は文書長の影響を受けやすいため，以下の式(6)に示す正規化手法を用いた．単語tの出現頻度を$tf_{req}(t,d)$，文書d に含まれる単語数を$t_{num}(d)$ とする．

また文章内の全ての索引語数で割った相対頻度を用いる．

研究報告用原稿の作成から投稿までの流れは，次の通りである．

$$tf(t,d) = \frac{\log(tf_{req}(t,d)+1)}{\log(t_{num}(d))} \quad (6)$$

またidfとは，ある索引語がどの程度その文章に特徴的に現れるのか，という特定性を表す尺度である．そのため，ある文章内だけではなく，文章群が存在する空間全体での索引語の分布を調べる必要がある．$idf(t)$は文書数Nと索引語tが出現する文書の数$idf\ (t)$によって決まり，式(7)によって定義される．

$$idf(t) = \log \frac{N}{df(t)} + 1 \quad (7)$$

ここで，Nは対象とする空間内にある文章の総数，$idf(t)$は索引語tが出現する文章数である．idfは，多数の文章に出現する語なら小さく，逆に特定の文章にしか出現しない語なら大きい値をとる．また，ここで対数をとるのはidfの変化を小さくするためである．つまりidfとは，ある索引語がその文章を特徴付けている程度を示す．

これら 2 つの情報，索引語頻度であるtfと索引語の特定性を表すidfを組み合わせたものを索引語の重みとする．この重みは，文章dにおける索引語tの重みを以下の式のようにtfとidfの積で定義しており，網羅性と特定性をともに考慮した重みである．

$$w_t^d = tf_d(t) \times idf(t) \quad (8)$$

本稿では，式(8)で得られた値を索引語tの重みとする．

2.4 南瓜

南瓜とは係り受け解析を行うシステムである．係り受けとは語の間にある修飾-被修飾の関係のことで，文の係り受け関係を句・文節を単位として解析を行うのが係り受け解析である．二つの文節の間に係り受け関係があるかどうかの判断ができる．また，現在の南瓜 0.68 というバージョンでは，茶筅[7]による形態素解析，単語を単位とするチャンキング[7]による固有表現同定，文節へのまとめ上げ，および文節間の係り受け解析を連続的に行う．

```
* 0 5D 1/2
五　　名詞,数,*,*,*,*,五,ゴ,ゴ
時　　名詞,接尾,助数詞,*,*,*,時,ジ,ジ
に　　助詞,格助詞,一般,*,*,*,に,ニ,ニ
* 1 2D 1/2
変圧　　名詞,サ変接続,*,*,*,*,変圧,ヘンアツ,ヘンアツ
器　　名詞,接尾,一般,*,*,*,器,キ,キ
の　　助詞,連体化,*,*,*,*,の,ノ,ノ
* 2 3D 0/1
漏電　　名詞,サ変接続,*,*,*,*,漏電,ロウデン,ローデン
の　　助詞,連体化,*,*,*,*,の,ノ,ノ
* 3 5D 0/0
ため　　名詞,非自立,副詞可能,*,*,*,ため,タメ,タメ
* 4 5D 0/1
障害　　名詞,一般,*,*,*,*,障害,ショウガイ,ショーガイ
が　　助詞,格助詞,一般,*,*,*,が,ガ,ガ
* 5 -1D 1/2
発生　　名詞,サ変接続,*,*,*,*,発生,ハッセイ,ハッセイ
し　　動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
た　　助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
```

図 1　形態素解析結果

図 1 に，南瓜の係り受け解析出力例を図示する，各行が一つの文節に対応し，それぞれがどの文節に係るか表現でき，“D”という記号によって示されている．この例では，「ため」，「障害が」，「五時に」が文末の「発生した」に係り，「ため」は「漏電の」に係り，「漏電の」は「変圧器の」に係ることなどがわかる．日本語の文節は，必ず前から後ろに係り，異なる係り受けは互いに交差しない（話し言葉などで例外はあるが）と仮定しているので，図 1 に示したような木構造で表示することができる．このような解析を行うことより，文中では「発生した」という動詞が「障害が」という名詞を直接修飾していることなどを知ることができる．

3. EMD を用いた文検索

文間の類似度を求める際，正しい計算順番をもとにうまく計算できないと文間の正確な類似度を求めることはできない．計算の仕方としては，様々な方法が考えられ，例えば単語間の関連性が高い順に単語の対応をとり計算する方法などが挙げられる．しかし 1 対 1 で対応をとる方法では，要求文と対象文の語の少ない方の語数でしか対応がとれない．例えば要求文の語が 3 語，対象文の語が 100 語であった場合，対象文の 97 語は計算の対象外となる．さらに実際の検索において，ユーザは要求文にあまり多くの語を入力しないため，要求文と対象文との語数の差は非常に大きくなると考えられる．そのため文の単語の重要性と単語間の関連性を考慮しM対Nで対応を取る必要がある．

そこで，本稿では類似画像検索の分野で注目されている EMD を用いて文間の類似度を算出する方法を用いる．EMD は輸送問題における輸送コストの最適解を求めるアルゴリズムであり，需要地及び供給地の重みと需要地と供給地間の距離を定義できれば文間の関連性としても適用できる．この EMD を用いることで単語の重みと単語間の関連性を考慮して柔軟に対応を取り，文間の関連度を求めることができる．

EMD とは 2 つの離散分布において，一方の分布を他方の分布に変換するための最小コストとして定義される．EMD を求める際，二つの分布は要素の重み付き集合として表現される．一方の分布 P を集合として表現すると$p = \{(p_1, w_{p_1}), \dots, (p_m, w_{p_m})\}$となる．今，分布$P$は$m$個の特徴量で表現されており，$p_i$ は特徴量，w_{p_i} はその特徴量に対する重みである．同様に，一方の分布Qも集合として表すと，$Q = \{(q_1, w_{q_1}), \dots, (q_n, w_{q_n})\}$となる．今，$p_i$と$q_i$の距離を$d_{ij}$とし，全特徴間の距離を$D = [d_{ij}]$とする．本研究においての距離は索引語間の関連性を表し，索引語間の関連性が高い場合に距離は近く，関連性が低い場合は遠くなる．本研究では距離は 1 から一致度の値を引いた値とした．ここで，p_iからq_iへの輸送量をf_{ij}とすると，全輸送量は$F = [f_{ij}]$となる．ここで，式(9)に示すコスト関数を最小とする輸送量F

を求め，EMD を計算する．

$$WORK(P,Q,F) = \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}f_{ij} \quad (9)$$

ただし，上記のコスト関数を最小化する際，以下の制約条件を満たす必要がある．

$f_{ij} \geq 0, \quad 1 \leq i \leq m, \ 1 \leq j \leq n$

$\sum_{j=1}^{n} f_{ij} \leq w_{p_i}, \ 1 \leq i \leq m \quad \sum_{i=1}^{m} f_{ij} \leq w_{q_j}, \ 1 \leq j \leq n$

$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = min\left(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j}\right)$

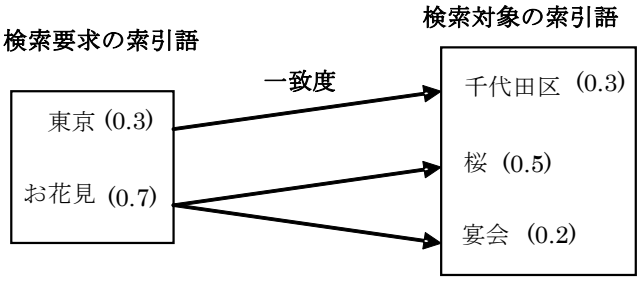
以上の制約条件の下で求められた最適な全輸送量Fを用いて分布P, Q間の EMD を以下のように求める．

$$EMD(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \quad (10)$$

ここで，最適なコスト関数$WORK(P,Q,F)$を EMD としてそのまま用いないのは，コスト関数は輸送元もしくは輸送先の重みの総和に依存するので，正規化することによってその影響を取り除くためである．P，Q間の EMD 関連度は$DoM_{EMD}(P,Q)$とする．

$$DoM_{EMD}(P,Q) = 1 - EMD(P,Q) \quad (11)$$

EMD の文検索への適用として要求文に『東京でお花見をする』，対象文に『千代田区の桜で宴会した』という二つの文に対して EMD を文検索に適用した例を図 2 に示す．なお，図 2 の括弧内の数字はその索引語の重みである．

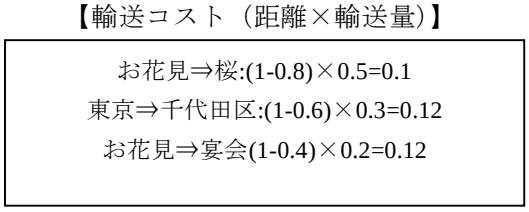


	千代田区	桜	宴会
東京	0.6	0.1	0.2
お花見	0.3	0.8	0.4

図 2　EMD を文検索に適用した例

EMD を文書検索に適用するには，各需要地と供給地間の距離を定義する必要がある．需要地としては，検索要求の索引語を，供給地としては検索対象の索引語を割り当てる．需要量としては要求文の索引語の重み，供給量としては対象文の索引語の重みを用いる．そして，需要地と供給地間の距離は索引語間の関連性と見立てることができ，概念ベースを用いた一致度により求めることができる．一致度は関連性が高いと値も大きくなるため，1 から一致度を引いた値を需要地と供給地間の距離とする．以上の値を用いれば EMD による定量化が可能になる．

どの(検索対象の)単語から優先的に重みを配分するかについて，最初に，検索要求の索引語の重みを検索対象の索引語へ一致度の高い順に配分する．課題の「東京」という単語において，検索対象の中で一致度最も高い組み合わせは「東京，千代田区」の 0.6 であるため，「千代田区」に「東京」の重みを搬送する．「東京」の重みが 0.3 であり，「千代田区」の容量が 0.3 であるため，「東京」という単語の重み配分は終了となる．「お花見」という単語において，検索対象の中で一致度最も高い組み合わせは「お花見，桜」の 0.8 であるため，「桜」に「お花見」の重みを搬送する．「お花見」の重みが 0.7 であり，「桜」の容量が 0.5 であるため，「お花見」の重みは 0.5 しか搬送できない．そのため，残る 0.2 を別の検索対象に搬送しなければならない．二番目に一致度の高い組み合わせは「お花見，宴会」の 0.4 である．これにより次の搬送対象は「宴会」となり，これに「お花見」の残りの重みを搬送する．「宴会」の容量が 0.2 であるため，残る 0.2 の搬送量が全部搬送できる．これにより，全ての検索要求の重み配分が完了する．この場合での重み配分後の関連度算出プロセスを図 3 に示す．



全体コスト：0.1+0.12+0.12=0.34
EMD：(0.1+0.12+0.12)÷1.0=0.34
文章関連度：1-0.34=0.66

図 3　EMD を用いた関連度算出プロセス

4. 依存構文

依存構文はルシアン・テニエールの構造統語論要説の理論に基づいている．テニエールのモデルは stemma と呼ばれ，語の間の文法的依存関係の図示に基づいている．文において動詞は最も高いレベルの語とみなされ，補語の集合の中心になる．また補語はさらにそれぞれの補語の中心になる．語を主部と述部に分ける考え方と異なり，テニエールの理論では主語も動詞に従属している．文中の語は，語の間の文法的依存関係に基づいた木構造で表現される．例として「漏電のため障害が発生した」という文を依存構文で分析した結果を図 4 に示す，この文においては，「発生した」を中心として，「ため・障害・漏電」が補語にあたる．

語を主部と述部に分ける考え方と異なり，テニエールの理論では主語も動詞に従属する．

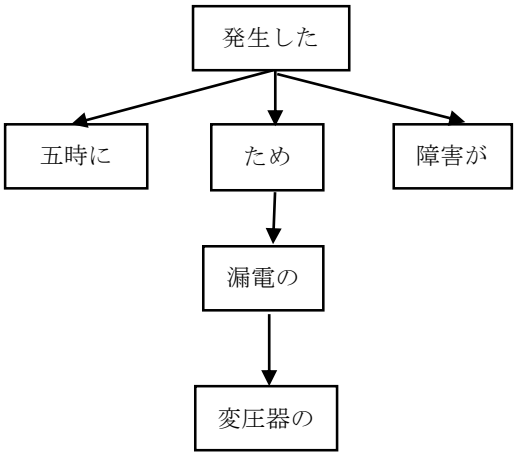


図 4　依存構文の例

依存構文によって一つの文の単語を重要単語と次要単語に分けて重要度をつけることで，単語の重要度を適切に表せると考えられる．重要単語は文の中で文の意味をもっと表せる単語と定義する．次要単語は重要単語以外の単語と定義する

南瓜で「漏電のため障害が発生した」という文を分析すると，図 5 のように結果が出力され，「D」によって単語の間の関係を表すことができる．

```
* 0 5D 1/2
五　　名詞,数,*,*,*,*,五,ゴ,ゴ
時　　名詞,接尾,助数詞,*,*,*,時,ジ,ジ
に　　助詞,格助詞,一般,*,*,*,に,ニ,ニ
* 1 2D 1/2
変圧　　名詞,サ変接続,*,*,*,*,変圧,ヘンアツ
器　　名詞,接尾,一般,*,*,*,器,キ,キ
の　　助詞,連体化,*,*,*,*,の,ノ,ノ
* 2 3D 0/1
漏電　　名詞,サ変接続,*,*,*,*,漏電,ロウデン
の　　助詞,連体化,*,*,*,*,の,ノ,ノ
* 3 5D 0/0
ため　　名詞,非自立,副詞可能,*,*,*,ため,タメ,タメ
* 4 5D 0/1
障害　　名詞,一般,*,*,*,*,障害,ショウガイ
が　　助詞,格助詞,一般,*,*,*,が,ガ,ガ
* 5 -1D 1/2
発生　　名詞,サ変接続,*,*,*,*,発生,ハッセイ
し　　動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
た　　助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
```

図 5　依存構文の分析例

図 5 の結果から，「五時に」，「ため」と「障害が」の三つは「発生した」に依存しており，また「漏電」は「ため」に依存しており，さらに「変圧器の」は「漏電の」に依存している．これらは依存構文の木構造になっているので，南瓜を用いて文を依存構文の木構造に分析できることが分かる．以上の例の場合は重要単語のグループは「五時に」，「ため」，「障害が」と「発生した」という四つの単語の集団になっている．次要単語のグループは「変圧器の」と「漏電の」という二つの単語の集団になっている．

5. 提案手法

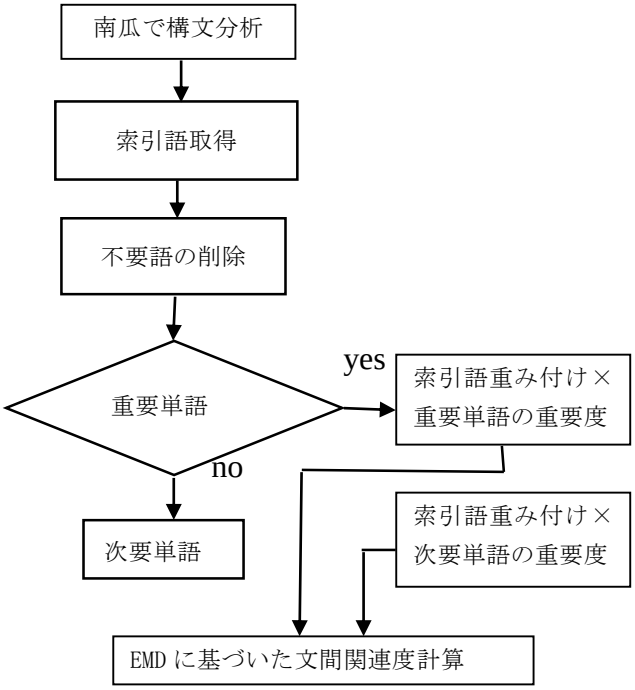


図 6　依存構文 EMD に基づいた文間の関連度計算の全体図

提案する依存構文とEMDに基づいた文間の関連度計算手法の全体図を以下の図6に示す．まず，二つの文章は南瓜で構文分析を行い，索引語を取る．本稿では日本語での検索を想定している．日本語は英語などとは異なり，単語間に明確な区切りがない．そこで，文章から単語を切り出す必要がある．そこで，形態素解析器である南瓜を用いて単語の切り出しを行い，それぞれの索引語に重み付けを行う．それから，不要語の削除をする．本研究では不要語を削除するために評価データの実験に使用する検索課題と検索対象の空間でのtf値とidf値をもとに検索対象内の不要語を削除した．不要語とは名詞，動詞，形容詞の以外の単語になる．また，名詞，動詞，形容詞の中で文書の意味を表すことができない頻度が高い単語になる．具体的に言えば，「し」，「いたし」，「あり」「ある」，「した」，「なった」，「いたした」，「い」，「ため」，「為」，「いら」，「-」，「,」，「したい」，「する」という単語のリストになる．分析した単語が重要単語かどうか判断し，重要単語の集団によって単語の重み付け（$tf \cdot idf$値）を計算する．そして，計算した重み付けは重要単語の重要度と掛け算を行って最終の単語重み付けになる．単語の重要度は単語の重要性を表す値と定義する．本研究は，単語の重要度は重要単語と次要単語の重要度という二つの

値になる．また，重要単語以外の単語は次要単語の集団になる，次要単語によって重要単語と同じ方法で単語の重み付けを計算する．

重要単語と次要単語の決定方法に対して依存構文に基づいて重要単語を木構造の中心の単語とそれに直接に繋がっている名詞，動詞，形容詞，副詞とする方法である．次要単語は重要単語以外の単語とする．これを重要単語と次要単語の決定とする．

また，本研究では形態素解析ソフト南瓜を利用し，「名詞・形容詞・動詞」を索引語として用いる．索引語に対する重み付けは$tf \cdot idf$を使用する．文章S_1, S_2の EMD を用いた関連度は$D_{EMD}(S_1, S_2)$と定義する．

文章を南瓜で構文分析して，木構造の中心単語とそれに直接繋がっている重要単語H,それ以外の索引語を次要単語Aとする．このとき，以下の式 (12)，(13) に示す正規化手法を用いた．単語tの重要単語の中で出現頻度を$tf_{reqh}(t,d)$，重要単語の数はn,単語fの次要単語の中で出現頻度を$tf_{reqh}(f,d)$，次要単語数はm,文書dに含まれる単語数を$t_{num}(d)$ とする．

重要単語の単語tのtf値をtf_hとする．

$$tf_h(\mathrm{t,d}) = \frac{\log(tf_{reqh}(t,d)+1)}{\log(t_{num}(d))} \times a_1 \quad (12)$$

次要単語の単語fのtf値をtf_aとする．

$$tf_a(\mathrm{t,d}) = \frac{\log(tf_{reqa}(f,d)+1)}{\log(t_{num}(d))} \times a_2 \quad (13)$$

ただし，a_1，a_2は定数値であり，さらに，$\sum_{i=0}^{n-1} tf_{hi} + \sum_{j=0}^{m-1} tf_{aj} = 1$になる．

文章S_1, S_2を南瓜で構文分析して，それぞれの重要単語をH_1, H_2，次要単語をA_1, A_2に分け，重要単語の索引語の重みはH_{tfidf1}，H_{tfidf2}，次要単語の索引語の重みはA_{tfidf1}, A_{tfidf2}とする．EMD を用いて関連度を求める．以下の式で表現される．

$$D_{EMD}(S_1, S_2) = D_{EMD}\left(H_{tfidf1}, A_{tfidf1}, H_{tfidf2}, A_{tfidf2}\right) \quad (14)$$

6. 評価方法

既存手法である EMD を用いた文間関連度計算方式と本稿で提案する文間関連度計算方式について精度に対して評価を行う．

本研究では提案手法の効果を検証するため，日本での代表的な情報検索システム評価用のテストコレクションであるNTCIR[9]を用いた．NTCIRは文献データ集合，検索課題集合，各検索課題に対する文献の適合不適合判定からなるもので，同一のテストコレクションを利用することにより共通の基準で情報検索システムを評価することができるようにしたものである．その中でも，本研究では一般の利用者が実際に検索する環境に近いWeb 検索用のテストコレクションであるNTCIR3-WEB を用いた．図7 にNTCIR3-WEB の検索課題の一例を示す．

```
<TOPIC>
<NUM>0010</NUM>
<TITLE CASE="b">オーロラ，条件，観測</TITLE>
<DESC>観測のために、オーロラの発生する条件が知りたい</DESC>
<NARR><BACK>オーロラを観測するために、発生に必要な条件や、発生のメカニズムが知りたい。
</BACK><RELE>オーロラ観測記などは、場所と日時が表記されており、発生時の天候・温度等を追跡調査することが可能な物のみ適合とする。</RELE></NARR>
<CONC>オーロラ，発生，条件，観測，メカニズム
</CONC>
<RDOC>NW003201843, NW001129327,
NW002699585</RDOC>
<USER>大学院修士1年，女性，検索歴2.5年</USER>
</TOPIC>
```

図7 NTCIR3-WEB評価セットの課題の例

検索課題には NUM, TITLE, DESC, NARR, CONC, RDOC, USER の 7 つのフィールドが含まれているが，標準的なフィールドは TITLE(title), DESC(description), NARR(narrative), CONC(concept)の 4 つである．TITLE とは検索課題の内容を簡単に表したタイトル，DESC は検索する内容を文で記述したもの，NARR は検索する内容の詳細な説明，CONC は検索する内容を表すキーワードである．本研究では検索要求を文章で入力するシステムの開発を想定し，DESC のみを使用する．

本稿の評価には情報検索システムテストコレクション NTCIR3-WEB より，検索課題 36 件と検索対象 1,000 件を用意する．また正解文章リストが存在し，各検索課題に対して，各文章が H（高適合），A（適合），B（部分的適合），C（不適合）の 4 段階の適合度が設定されている．本稿では H 判定と A 判定を正解文章とする．

各検索課題に対して 1,000 件の検索対象全ての文間関連度を求め，文間関連度順に並べ替える．そして，正解文章リストを参照し適合文章の順位を調べ評価する．

評価指標には各検索課題の平均精度(Average Precision, AP)，平均の精度の平均(Mean Average Precision, MAP)を使用する．検索課題に対する平均精度 AP は，式(15)で定義される．まず順位 i 位の文章が適合しているならば 1，そうでなければ 0 となる変数をz_iとする．S を適合文章の総数，n は出力文章数である．

$$AP = \frac{1}{S}\sum_{i=1}^{n}\frac{z_i}{i}\left(1 + \sum_{k=1}^{i-1} z_k\right) \quad (15)$$

平均精度の平均(MAP)は，全ての検索課題に対して平均精度を平均したものであり，式(16)で定義される．具体的には，検索課題が T 件ありそれぞれの課題に対するあるシステムのAP_h(h=1,…,T)と表記すれば，その平均が MAP に相当する．

$$MAP = \frac{1}{T}\sum_{h=1}^{T} AP_h \tag{16}$$

以下の表 2 に例を示す.

表 2 平均精度の例

順位	適合･不適合	精度
1	○	1
2	×	0.5
3	○	0.67
4	○	0.75
5	×	0.6
6	○	0.67
7	×	0.57

平均精度の平均(Mean Average precision , MAP)は，平均精度を合計し検索課題数 36 で割ったものとなり，例の場合は MAP=（1+0.67+0.75+0.67）/4=0.7725 となる.

また，本実験において索引語の重み付けである$tf \cdot idf$のtf値の空間dは評価セットの各検索課題（一文）或は検索対象（一文章）である．idf値に対する文章空間Nは評価セットの検索課題 36 件と検索対象 1000 件である.

7. 評価結果

5 章に示した提案手法について，重要単語×ｎ，次要単語×(1－ｎ)， ただし，0.0≤ｎ≤1.0 というように重要単語と次要単語の重み配分を変えて評価実験を行った結果を示す.

7.1 提案手法の重要単語と次要単語の決定の評価結果

提案手法の重要単語と次要単語の決定 (2) における平均精度の平均（MAP）を$n = 0.0$から$n = 1.0$まで変化させた時の結果を図 8 に示す.

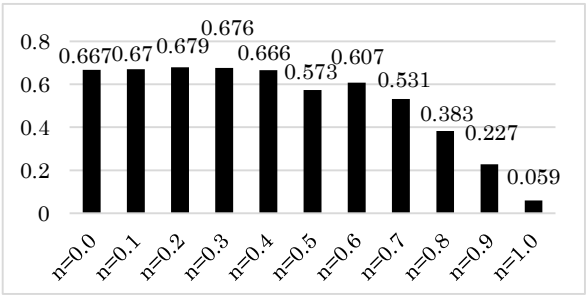


図 8　提案手法の重要単語と次要単語の決定における MAP

図 8 に示す結果として精度はnの減少とともに，上がっていることが分かる．原因として，今回の評価セットの検索課題は「○○を知りたい」という形の文が多いことが考えられる．たとえば，検索課題：「サルサを踊れるようになる方法が知りたい」を南瓜で分析すると，重要単語は「知り」,「方法」という単語となる．また次要単語は「サルサ」,「踊れる」,「よう」,「なる」という単語となる．この文において，重要単語の「知り」,「方法」は文の意味を適切に表せていないと考えられる．逆に次要単語の「サルサ」,「踊れる」,「よう」,「なる」は文の意味が表せると考えられる．これより，今回の評価セットに対しては次要単語のほうが文の意味を適切に表していると考えられる．また，n=0.2 のときから n=0.0 のときまで，ｎ＝0.2 のとき精度が頂点になって，それからほぼ変わらない．ｎ＝0.2 のときから，重要単語は精度に影響が少ないと考えられる.

7.2 結論

また構文分析によっての提案手法において次要単語の重み付けを増加したら,既存手法より平均精度の平均(MAP)は高くなることが分かった．特に提案手法の重要単語と次要単語の決定で重要単語×0.2・次要単語×0.8 にした時,最も高い MAP が出てきたので，構文分析によっての提案手法において重要単語と次要単語の決定で重要単語×0.2･次要単語×0.8 にすべきであると考えられる.

8. おわりに

本稿では索引語の関連性を概念ベースにより定量化し，それをもとに文中の単語を，文の構造から重要単語と次要単語に分けて重みをつける考え方と，重要単語と次要単語を分けて重みをつける考え方の方法により文章の関連性の定量化手法の精度向上を目指した．その評価を Web 検索評価用テストコレクション NTCIR を用いて検証した．結果として検索精度を向上することができた．構文分析理論においては重要単語よりも次要単語の重みの配分を大きくする方が精度が高くなるという結論となったため，実際の検索システムに応用していくことが考えられる．また，今回の検索対象（文章）は句点によって文を分けたが，日本語では複文や重文のように 1 つの文の中に 2 つ以上の主述関係を持つ文も存在する，たとえば,「物事を多角的な角度から考える力学校でを学んだ、そして先輩・後輩関係を含めた厳しい上下関係も学んだ.」という文があれば，句点によって一つの文として見えるが,「そして」前の文と後ろの文の意味が違う.「そして」前の文と後ろの文は二つ文として分析した方が意味的に正しいと考えられる．そのため,「そして」など接続詞を含める文に対して適切に文を分けること

が必要になると考えられる．

謝辞

本研究の一部は，科学研究費補助金（若手研究（Ｂ）24700215）の補助を受けて行った．

参考文献

[1] ウィキペディア“Google” https://ja.wikipedia.org/wiki/Google，（参照 2016-01-10）
[2] 藤江悠五，渡部広一，河岡司，“概念ベースと Earth Mover’s Distance を用いた文書検索”，自然言語処理，Vol.16，No.3，p325-349，2009．
[3] G.Salton,C.Buckley,Term-weight in gapproaches in automatic text retrieval,Information Processing and Management,Vol.41, No．4,pp.513-523,1988．
[4] ルシアン・テニエール（著），小泉 保（翻訳），“構造統語論要説”，研究社．2007 年．ISBN 4327401455．
[5] 奥村紀之，土屋誠司，渡部広一，河岡司，“概念間の関連度計算のための大規模概念ベースの構築”，自然言語処理，Vol.14，No.5，pp.41-64，2007．
[6] 渡部広一，奥村紀之，河岡司，“概念の意味属性と共起情報を用いた関連度計算方式”,自然言語処理,Vol.13，No．1，pp．53-74，2006．
[7] 松本裕治，高岡一馬，浅原正幸，工藤拓“茶筌と南瓜による日本語解析”， http://chasen.naist.jp/hiki/ChaSen/,（参照 2015-11-10）
[8] WU Zuoyan，WANG Yu．New measure of sentences similarity based on hierarchical network of concepts theory and dependency parsing．Computer Engineering and Applications，Vol.50, No.3, pp．97—102, 2014．
[9]NTICRProject“NTCIR|HOME”http://research.nii.ac.jp/ntcir/index-ja.html，（参照 2015-12-10）