目的

- 専門用語の抽出
  - サジェスチョン
  - メタ情報
  - クラスタリング

専門用語の抽出

・コーパスから固有表現を自動抽出する
・対象：単名詞とその単名詞からなる複合名詞のみとする

・固有表現
  ・単名詞
    →これ以上分割できない名詞
  ・複合名詞
    →専門用語の多くは複合語
  ・ターム性
    ・ある単語単位(複合名詞など)のもつ専門用語固有概念への関連の強さ
    ・専門用語を書いた専門家の概念に直結していると考えられる
    ・tf-idfは表層表現のコーパスでの現れ方を利用した近似表現に過ぎない
      →ターム性を直接的に反映する用語抽出方法が必要
  ・ユニット性
    ・ある言語単位がコーパス中で使用されている度合い

単名詞の連接統計情報の一般化

・特定のコーパスを想定したとき、単名詞Nが連接する状況
すなわち単名詞バイグラム(2-gram)を以下のように表す

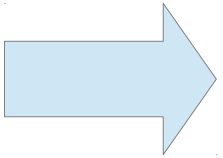
$$
\begin{array}{ll}
[LN_1 \quad N](\#L_1) & [N \quad RN_1](\#R_1) \\
[LN_2 \quad N](\#L_2) & [N \quad RN_2](\#R_2) \\
\vdots & \vdots \\
[LN_n \quad N](\#L_n) & [N \quad RN_m](\#R_m)
\end{array}
$$

図 1: 単名詞 N を含む単名詞バイグラムと左右連接単名詞の頻度

・LN_i(i=1,...,n)は、単名詞バイアグラム[LN_i　N]においてNの左方に連接する単名詞(n種類)を表す
・RN_i(i=1,...,m)は、単名詞バイアグラム[N　RN_i]においてNの右方に連接する単名詞RN_i(m種類)を表す
・#L_i(i=1,...,n)は、LN_iの頻度
・#R_i(i=1,...,m)は、RN_iの頻度

実例（頻度と種類）

トライグラム　統計
トライグラム
単語　トライグラム
クラス　トライグラム
単語　トライグラム
トライグラム
トライグラム　抽出
単語　トライグラム　統計
トライグラム
文字　トライグラム



[単語　トライグラム](3)
[クラス　トライグラム](1)
[文字　トライグラム](1)
#LDN(トライグラム)=3
#LN(トライグラム)=5

[トライグラム　統計](2)
[トライグラム　抽出](1)
#RDN(トライグラム)=2
#RN(トライグラム)=3

・連接種類数
- #LDN(N) : 単名詞バイグラムで単名詞Nの左方にくる単名詞の種類の異なり数
- #RDN(N) :　〃　右方　〃

・連接頻度
- #LN(N) : Nの左方に連結して複合名詞を形成する全単名詞の頻度
- #RN(N) : 〃　右方　〃

専門用語には複合名詞が多い

複合名詞のスコアを定義する必要がある

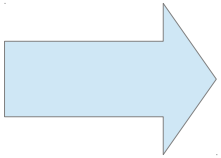
複合名詞のスコアリング

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

単名詞 : N_i
複合名詞 : CN = N_1N_2...N_L
単名詞Nの左方スコア関数 : FL(N)
単名詞Nの右方スコア関数 : FR(N)

実際に計算してみる(CN=トライグラム、L=1)

	トライグラム	統計
	トライグラム	
単語	トライグラム	
クラス	トライグラム	
単語	トライグラム	
	トライグラム	
	トライグラム	抽出
単語	トライグラム	統計
	トライグラム	
文字	トライグラム	



[単語　トライグラム](3)
[クラス　トライグラム](1)
[文字　トライグラム](1)
#LDN(トライグラム)=3
#LN(トライグラム)=5

[トライグラム　統計](2)
[トライグラム　抽出](1)
#RDN(トライグラム)=2
#RN(トライグラム)=3

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

連接頻度LN,RNを単名詞のスコアとした場合

LR(トライグラム)=$\sqrt{(5+1)(3+1)}$≒4.9

候補語の出現頻度を考慮した重み付け

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

f(CN) : 候補語CNが単独で出現した頻度

連結頻度をスコアとした場合

FLR(トライグラム) = f(トライグラム) × LR(トライグラム)
= 3 × $\sqrt{(5+1)(3+1)}$
= 14.90

C-Value

比較のために、単名詞バイグラムによらない用語スコアリングとして
C-Value(Frantzi and Ananiadou 1996)を考える

$$\text{C-value}(CN) = (length(CN) - 1) \times (n(CN) - \frac{t(CN)}{c(CN)})$$

CN:	複合名詞[1]
$length(CN)$:	CN の長さ (構成単名詞数)
$n(CN)$:	コーパスにおける CN の出現回数
$t(CN)$:	CN を含むより長い複合名詞の出現回数
$c(CN)$:	CN を含むより長い複合名詞の異なり数

length(CN)=1 (CNが単名詞)の場合、C-Valueが0になってしまう

MC-Value (Modified C-Value)

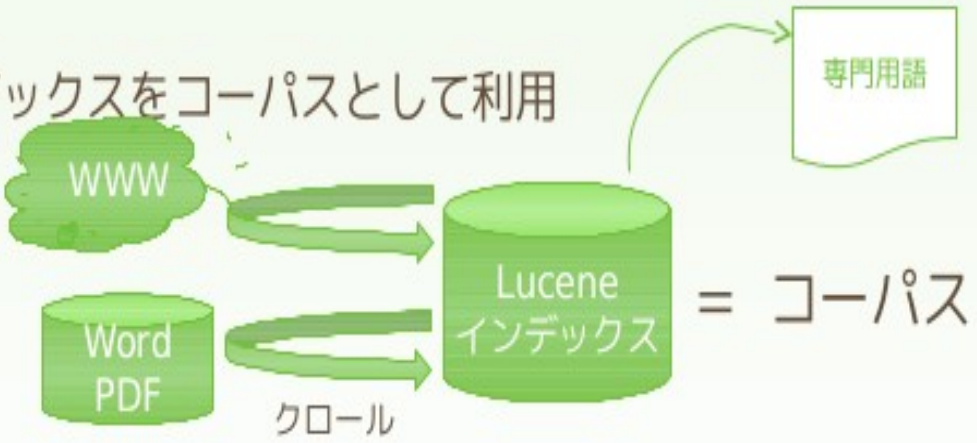
$$\text{MC-value}(CN) = length(CN) \times \left(n(CN) - \frac{t(CN)}{c(CN)}\right)$$

◯CN = トライグラム　の場合
length(トライグラム) = 1　: トライグラムの長さ(構成単名詞数)
n(トライグラム) = 10　　: コーパスにおけるトライグラムの出現回数
t(トライグラム) = 7　　: トライグラムを含む、より長い複合名詞の出現回数
c(トライグラム) = 6　　: トライグラムを含む、より長い複合名詞の異なり数

MC-Value(トライグラム) = 10 – 7/6 ≒ 8.83

実際に抽出してみる

- Luceneインデックスをコーパスとして利用



「出現頻度と連接頻度に基づく専門用語抽出」

http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/academic-res/jnlp10-1.pdf
「Luceneインデックスからの専門用語抽出」
http://www.slideshare.net/KojiSekiguchi/lucene-terms-extraction