Automatic Discovery of Similar Words

by

Pierre P. Senellart and Vincent D. Blondel

新納浩幸

Overview

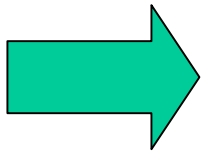
類義語をコーパス、Web、辞書などから集める手法

代表的3手法（リソースはコーパス）

（1）文書ベクトルモデル
（2）低頻出語のシソーラス
（3）SEXTANT

Web をリソースとした手法

} サーベイ



辞書から類義語を取り出す手法を提案

グラフ理論的な手法

背景

目的は、情報検索のクエリ拡張、シソーラスの自動作成、類義語辞書の作成

難しいタスク、反意語と類義語は似ている

類似の定義が曖昧

シソーラスの作成と同じ手法が使えるが、タスクとしては異なる。シソーラスの作成は語彙が固定される。

コーパスからの類義語の発見

シソーラスの自動構築を目的に研究されてきた

作成されるシソーラスは domain specific、
追加・修正が容易という長所がある

代表的3つの手法を紹介

（1）文書ベクトルモデル
（2）低頻出語のシソーラス
（3）SEXTANT

Document Vector Space Model

単語 → n 次元ベクトル

$$w \longrightarrow (f_1, f_2, \cdots, f_n)$$

$$f_i = \begin{cases} 1 & w \text{ が文書 i に含まれている} \\ 0 & w \text{ が文書 i に含まれていない} \end{cases}$$

$$sim(\vec{w}_1, \vec{w}_2) = \cos(\vec{w}_1, \vec{w}_2) = \frac{\vec{w}_1 \cdot \vec{w}_2}{|\vec{w}_1||\vec{w}_2|}$$

文書が直交していないと理論的にはおかしい

Cluster Measure

$$sim(\vec{w}_1, \vec{w}_2) = cluster(\vec{w}_1, \vec{w}_2) = \frac{\vec{w}_1 \cdot \vec{w}_2}{\| \vec{w}_1 \|_{l_1}}$$

H. Chen and K.J. Lynch.
Automatic construction of networks of concepts characterizing.
IEEE Transactions on Systems, Man and Cybernetics, 22(5): 885-902, 1992.

例）

$$a = (0,2,1,-1,0) \Rightarrow$$

$$\| a \|_{l_1} = |0| + |2| + |1| + |-1| + |0| = 4$$

Cos と Cluster の比較

再現率

Cluster Measure ＞ Cos measure , Human

適合率

Cluster Measure ≈ Cos measure ＜ Human

H. Chen and K.J. Lynch.
Automatic construction of networks of concepts characterizing.
IEEE Transactions on Systems, Man and Cybernetics, 22(5): 885-902, 1992.

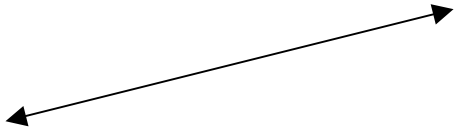
低頻出語のシソーラス

C.J. Crouch.
An approach to the automatic construction of global thesauri.
Information Processing and Management, 26, 629-640, 1990.

情報検索のクエリを修正する目的で、
低頻度語のシソーラスを作成

（1） 文書をベクトル空間モデルで表現し、
文書をクラスタリングする

（2） 各クラスターから indifferent discriminator を取り出す



（3） 低頻度語のシソーラスが作成される

Term discrimination value

単語を次元としたベクトル空間モデルで文書間距離を測定

単語 W を次元に加える

- 距離が増す → term discrimination value は高い
- 距離が減る → term discrimination value は低い

indifferent discriminator

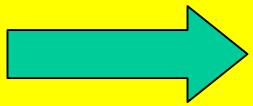
文書間の距離をほとんど変えない単語

TDV の設定方法

Term Discrimination Value（TDV）は現実には測定困難

頻度で設定する

文書数 1% 以下の出現単語



indifferent discriminator

１%～１０%： TDV の値は高い
頻度の高い単語: TDVの値は低い

SEXTANT

Semantic EXtraction Text via Analyzed Networks of Terms

G. Grefenstette.
Automatic thesaurus generation from raw text using knowledge-poor techniques.
Making Sense of Words. 9th Annual Conference of the UW Center for
the New OED and Text Research, 9, 1993.

G. Grefenstette.
Explorations in Automatic Thesaurus Discovery.
Kluwer Academic, Boston, 1994.

部分的に構文解析を行って類義語を取り出す

（詳細は省略）

Web の利用

コーパスを対象とした手法は使えない、巨大すぎる

P.D. Turney.
Mining the Web for synonyms: PMI-IR versus LSA on TOEFL.
European Conference on Machine Learning, pp.491-502, 2001.

英語の試験 TOEFL と ESL を利用

検索エンジン Altavista を利用

単語 A に対する単語 B の類似度の score を4つ提案

４つの Score（１）

$$score_1(a,b) = \frac{hits(a \wedge b)}{hits(b)}$$

$$score_2(a,b) = \frac{hits(a \quad near \quad b)}{hits(b)}$$

４つの Score（２）

$$score_3(a,b) = \frac{hits((a \quad near \quad b) \wedge \neg((a \vee b) \quad near \quad "not"))}{hits(b \wedge \neg(b \quad near \quad "not")}$$

$$score_4(a,b)$$

$$= \frac{hits((a \quad near \quad b) \wedge Context \wedge \neg((a \vee b) \quad near \quad "not"))}{hits(b \wedge Context \wedge \neg(b \quad near \quad "not")}$$

Context に英語試験の質問を利用する

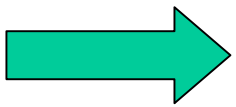
辞書からの類義語発見

本論文の中心

辞書の定義文から単語 w に対して隣接グラフ G_w を作成

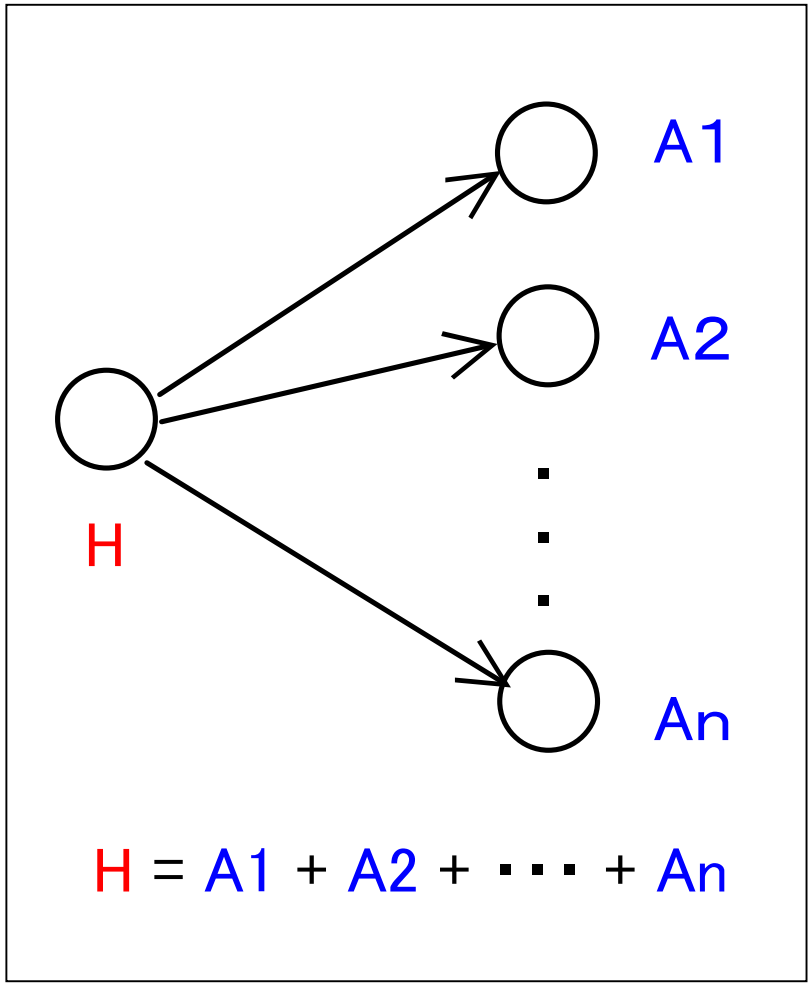
＊単語が頂点
＊単語の定義文に出てくる単語に有効辺がつく
＊単語 w を定義文に含む単語と、単語 w の定義文内の単語に限定した部分グラフ

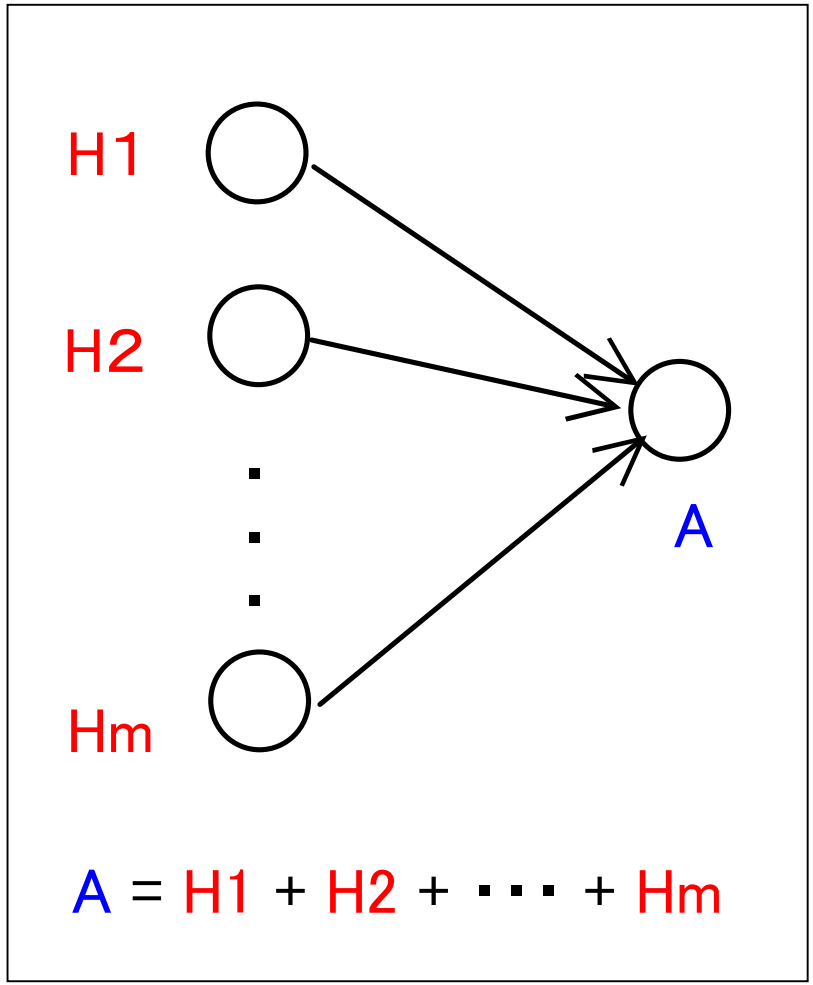
グラフの構造
「Hub と Authority の解析」に似た手法による単語へのスコア



単語 w と類義を発見する

Hub と Authority（1）





Hub と Authority（２）

グラフの頂点には Hub のスコア h と
Authority のスコア a がある

$$\begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} 0 & M \\ M^T & 0 \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

K+1 ステップ　　K ステップ

M: 隣接行列 (n 行 n 列)

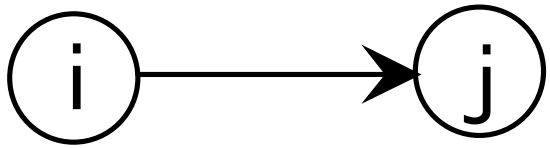
$$\vec{h} = eigen(MM^T)$$

$$\vec{a} = eigen(M^T M)$$

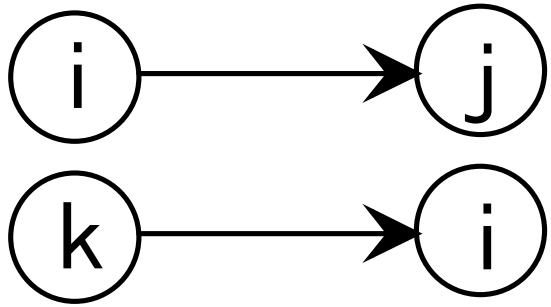
3つのスコア

G_w　上の各頂点に3つのスコアを考える

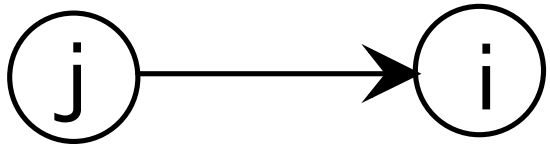
$$x_i^1 = \sum_j x_j^2$$



$$x_i^2 = \sum_j x_j^3 + \sum_k x_k^1$$



$$x_i^3 = \sum_j x_j^2$$



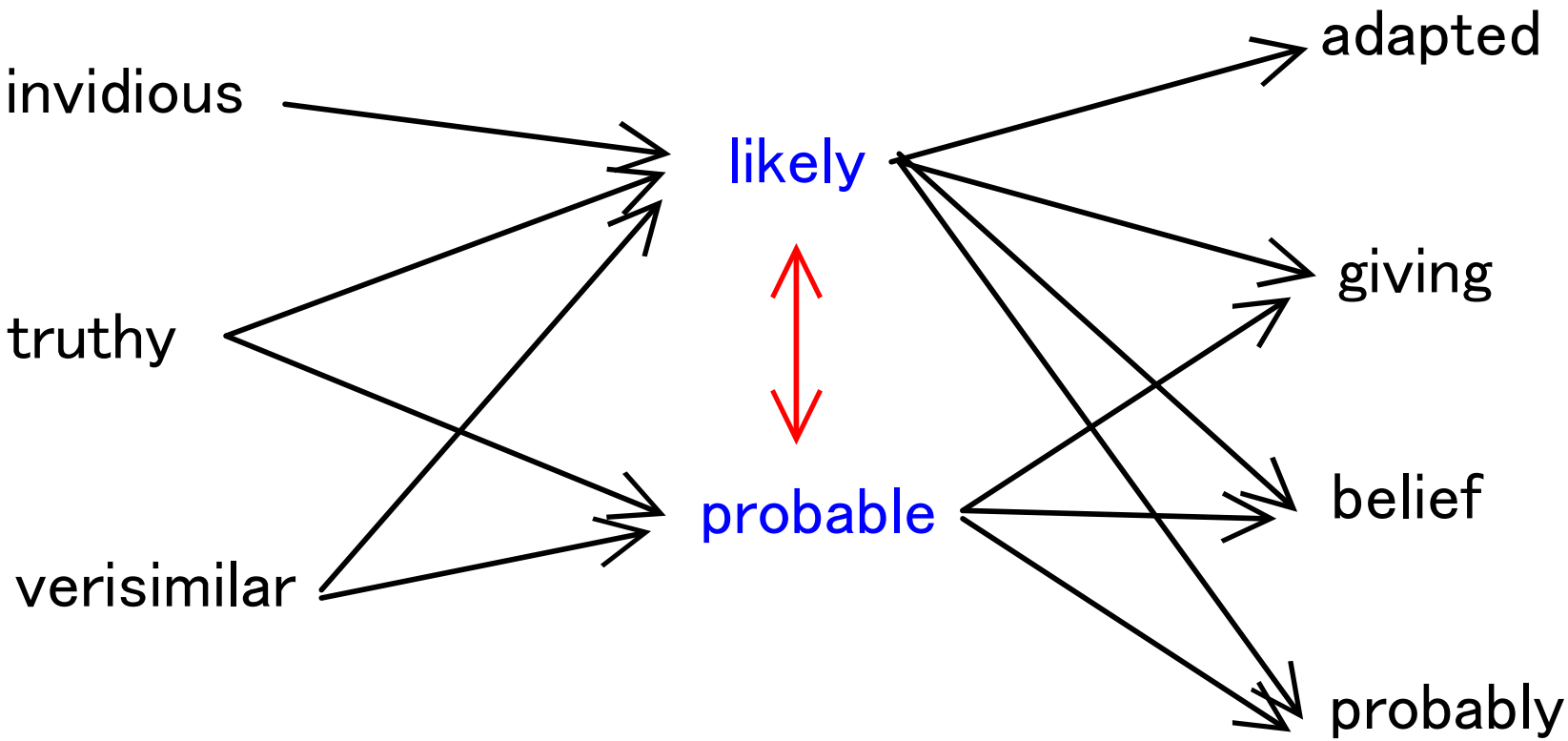
部分グラフ上の類義語

2の頂点が類義語の候補

その x^2 の値が類義語のスコア

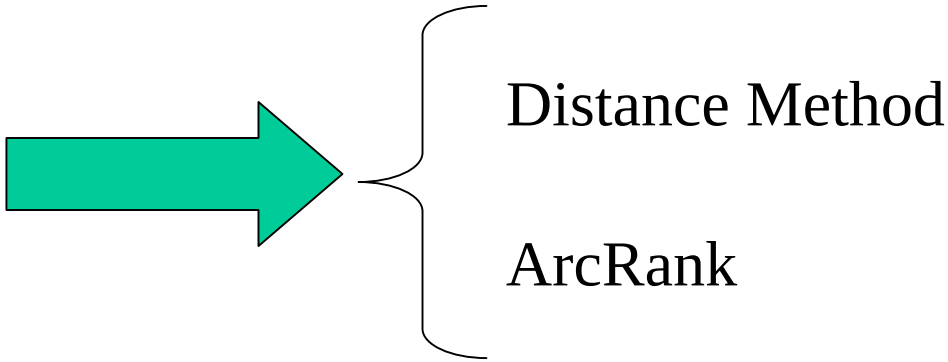
$$\overrightarrow{x^2} = eigen(MM^T + M^T M)$$

例


invidious
truthy
verisimilar
likely
probable
adapted
giving
belief
probably

その他のスコア

グラフ上の頂点間に類似度のスコアを与えられれば、
類義語が抽出できる



Distance Method

$$d(i,j) = \| A_{i,\bullet} - A_{j,\bullet} \|_1 + \| (A_{\bullet,i} - A_{\bullet,j})^T \|_1$$

注意

この設定では隣接グラフを取る必要はないが、
隣接グラフの方がよい結果が得られる

ArcRank

$$r(s,t) = \frac{p_s / |a_s|}{p_t}$$

$|a_s|$ ：s から出ているエッジの数

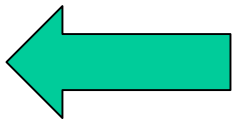
p_x ：頂点 x のページランク

グラフの作成方法

基本は

＊単語が頂点
＊単語の定義文に出てくる単語に有効辺がつく

実際の構築には、細かい処理が必要



説明省略、p.36　2.3.4

語尾変化の形やら、ストップワードの設定など

実験結果（disapper）

disappear　多くの類義語が存在

	Distance	Our method	ArcRank	Wordnet
1	**vanish**	**vanish**	**epidemic**	**vanish**
2	**wear**	**pass**	**disappearing**	**go away**
3	**die**	**die**	**port**	**end**
4	**sail**	**wear**	**dissipate**	**finish**
5	**faint**	**faint**	**cease**	**terminate**
6	**light**	**fade**	**eat**	**cease**
7	**port**	**sail**	**gradually**	
8	**absorb**	**light**	**instrumental**	
9	**appear**	**dissipate**	**darkness**	
10	**cease**	**cease**	**efface**	
Mark	3.6	6.3	1.2	7.5
Std dev.	1.8	1.7	1.2	1.4

○　　◎

実験結果（parallelogram）

parallelogram　限定された意味の単語、確かな類義語はない

（平行四辺形）

	Distance	Our method	ArcRank	Wordnet
1	**square**	**square**	**quadrilateral**	**quadrilateral**
2	**parallel**	**rhomb**	**gnomon**	**quadrangle**
3	**rhomb**	**parallel**	**right-lined**	**tetragon**
4	**prism**	**figure**	**rectangle**	
5	**figure**	**prism**	**consequently**	
6	**equal**	**equal**	**parallelepiped**	
7	**quadrilateral**	**opposite**	**parallel**	
8	**opposite**	**angles**	**cylinder**	
9	**altitude**	**quadrilateral**	**popular**	
10	**parallelepiped**	**rectangle**	**prism**	
Mark	4.6	4.8	3.3	6.3
Std dev.	2.7	2.5	2.2	2.5

△　△　×　×

実験結果（sugar）

sugar　多くの意味を持つ一般的な語

	Distance	Our method	ArcRank	Wordnet
1	**juice**	**cane**	**granulation**	**sweetening**
2	**starch**	**starch**	**shrub**	**sweetener**
2	**cane**	**sucrose**	**sucrose**	**carbohydrate**
4	**milk**	**milk**	**preserve**	**saccharide**
5	**molasses**	**sweet**	**honeyed**	**organic compound**
6	**sucrose**	**dextrose**	**property**	**saccarify**
7	**wax**	**molasses**	**sorghum**	**sweeten**
8	**root**	**juice**	**grocer**	**dulcify**
9	**crystalline**	**glucose**	**acetate**	**edulcorate**
10	**confection**	**lactose**	**saccharine**	**dulcorate**
Mark	3.9	6.3	4.3	6.2
Std dev.	2.0	2.4	2.3	2.9

○　◎　×

実験結果（science）

science　一般的な語だが曖昧、類義語は想定しづらい

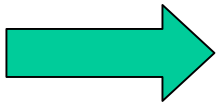
	Distance	Our method	ArcRank	Wordnet
1	art	art	formulate	knowledge domain
2	branch	branch	arithmetic	knowledge base
3	nature	law	systematize	discipline
4	law	study	scientific	subject
5	knowledge	practice	knowledge	subject area
6	principle	natural	geometry	subject field
7	life	knowledge	philosophical	field
8	natural	learning	learning	field of study
9	electricity	theory	expertness	ability
10	biology	principle	mathematics	power
Mark	3.6	4.4	3.2	7.1
Std dev.	2.0	2.5	2.9	2.6

？　？　△

Future Works

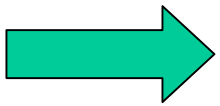
もっと大きなグラフでやる

例）cow の定義の中に ox は入っていない



単語に依存

辞書の種類を変更



他分野、他言語