# Web上の概念構造とその概要を用いた語義識別知識の獲得

佐々木稔

# はじめに

- 語義の識別
  - 対象単語の前後の文脈を利用
  - 文脈の捉え方：出現単語の類似性
- 問題点
  - 概念は同じでも使う単語が異なる場合がある
- 異表記同義語への対策
  - 辞書などの事前知識の活用

# Webからの事前知識獲得

- Wikipedia の利用
  - 「曖昧さ回避のためのページ」の利用
  - ラベル（個別の意味）とその説明の両方を利用
- 従来研究
  - Simone Paolo Ponzetto, Michael Strube: “Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution”,
    - Wikipedia の「曖昧さ回避のためのページ」のみ利用
  - Michael Strube, Simone Paolo Ponzetto: “WikiRelate! Computing Semantic Relatedness Using Wikipedia”,
    - Wikipedia のカテゴリのみを利用した単語類似度の計算手法

# 例:「核」の曖昧さ回避ページ

## 核

出典: フリー百科事典『ウィキペディア(Wikipedia)』

**核**(かく、さね)は、軸・中心となるもの。中核、核心。

英語のnucleus(形容詞 nuclear)、core、kernelなどの訳語にも使われる。

特に記したもの以外は「かく」と読む。

### 原子核 [編集]

- 原子核
- 核エネルギー
- 核兵器

### その他 [編集]

- 数学
  - 核 (数学)(カーネル)。準同型の核。
  - 積分変換の核(熱核、再生核)
- 自然科学
  - 核 (天体)。天体の中心部。
  - 細胞核。
  - 神経核。中枢神経において、神経細胞体の集合、細胞群。
  - 相転移の開始点となるもの。結晶化の場合は結晶核という。
  - 環式有機化合物の骨格部分。ベンゼン環(ベンゼン核)など。
  - 内果皮が硬化し、種子本体を保護するようになったもの。「さね」とも。
- 技術
  - カーネル。オペレーティングシステム(OS)の基本部分。
  - 真珠の養殖のときに母貝に入れる球。
- 社会
  - 父・母・1児のうち2要素ないし全てからなる家族を、核家族とよぶ。

### 固有名詞 [編集]

- 核 (CORE)。尾崎豊のシングルCD。

# 項目「原子核」の説明文

ウィキペディア
フリー百科事典

案内

- メインページ
- コミュニティ・ポータル
- 最近の出来事
- 新しいページ
- 最近更新したページ
- おまかせ表示
- 練習用ページ
- アップロード（ウィキメディア・コモンズ）

ヘルプ

- ヘルプ
- 井戸端
- お知らせ
- バグの報告
- 寄付
- ウィキペディアに関するお問い合わせ

検索

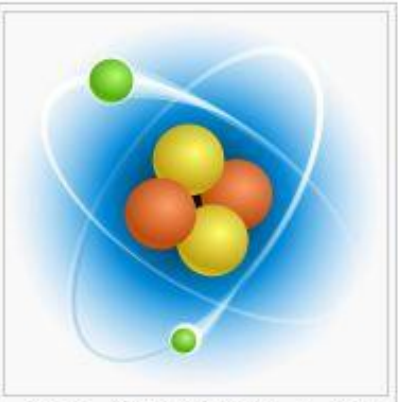
表示　検索

## 原子核

出典: フリー百科事典『ウィキペディア（Wikipedia）』

**原子核**（げんしかく）は、単に**核**（かく）ともいい、原子を電子と共に構成している。原子の中心に位置し核子の塊であり、正電荷を帯びている。核子は、通常の水素原子（軽水素）では陽子1個のみ、その他の原子では陽子と中性子から成る。

原子と比べて原子核は非常に小さく、たとえば最も小さい水素の原子核（つまり陽子）の大きさはおよそ半径 $10^{-15}$ m = 1 fmである。より重い原子核ではその質量数のほぼ1/3乗に比例して大きな半径を持つが、大きなもの、たとえば鉛でも10 fm を下回る。水素原子核以外では、その狭い空間に正電荷をもった陽子が複数存在するため、互いに大きな斥力（電磁気力）を受ける。この斥力に打ち勝って原子核を安定に存在させているのは、中性子の作用である。陽子、中性子の核子間には中間子を媒介した核力が引力として働き、これが電磁気的反発力に打ち勝って原子核を安定化させている。



ヘリウム原子の模式図。中心部の4つの球体からなる塊が原子核。周りを回っているのは電子である。

原子核の質量を半経験的に説明する、ヴァイツゼッカー＝ベーテ（Weizsaecker-Bethe）の半経験的（semiempirical）質量公式（原子核質量公式、他により改良された公式が存在する）がある。

原子核の安定性は、陽子、中性子の数と深く関わっており、特に原子核を安定にさせる数（魔法数）が存在する（液滴モデル、集団運動モデル、など）。ただし、最近の不安定核の研究によって極端に中性子過剰な核などではこれまで知られてきた魔法数の系列が消失することがわかってきた。

# Wikipediaを利用した事前知識獲得

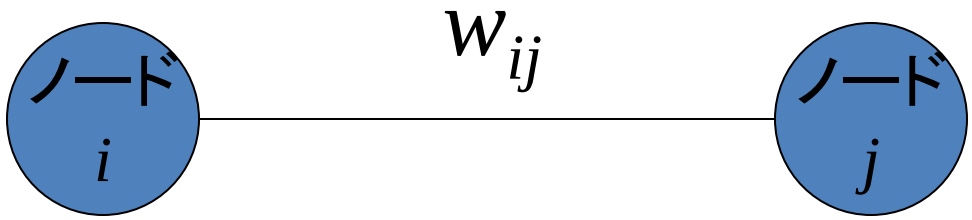
- 獲得手順
  1. 意味ラベルをルートハブとする
  2. 各意味の説明文に対して出現単語を抽出
  3. 出現単語の頻度を計算
  4. 頻繁に出現する単語を抽出
  5. ルートハブに頻出単語を接続する
  6. ノード間のエッジに頻度による類似度を付与
- 獲得手段
  - HyperLexアルゴリズムの最小スパン木を利用

# HyperLexアルゴリズム

- HyperLexによる語義識別（Agirre,Soroa 2007）
  - 単語共起によるグラフの作成
    - ノードは単語、エッジには相対頻度
  - 高密度なノード（ハブ）を求める
    - HyperLexアルゴリズム
  - 最小スパニング木を求める
  - 各文に対してスコアベクトルを計算
  - Markov Clustering (MCL) でクラスタリング

# 共起グラフの作成

- 文中の名詞に対し、共起グラフを作成
  - ノード：対象単語以外の名詞
  - エッジ：同じパラグラフに2単語あれば共起
  - エッジの重み $w_{ij}$ ：
    - ノード $i$ とノード $j$ の共起頻度



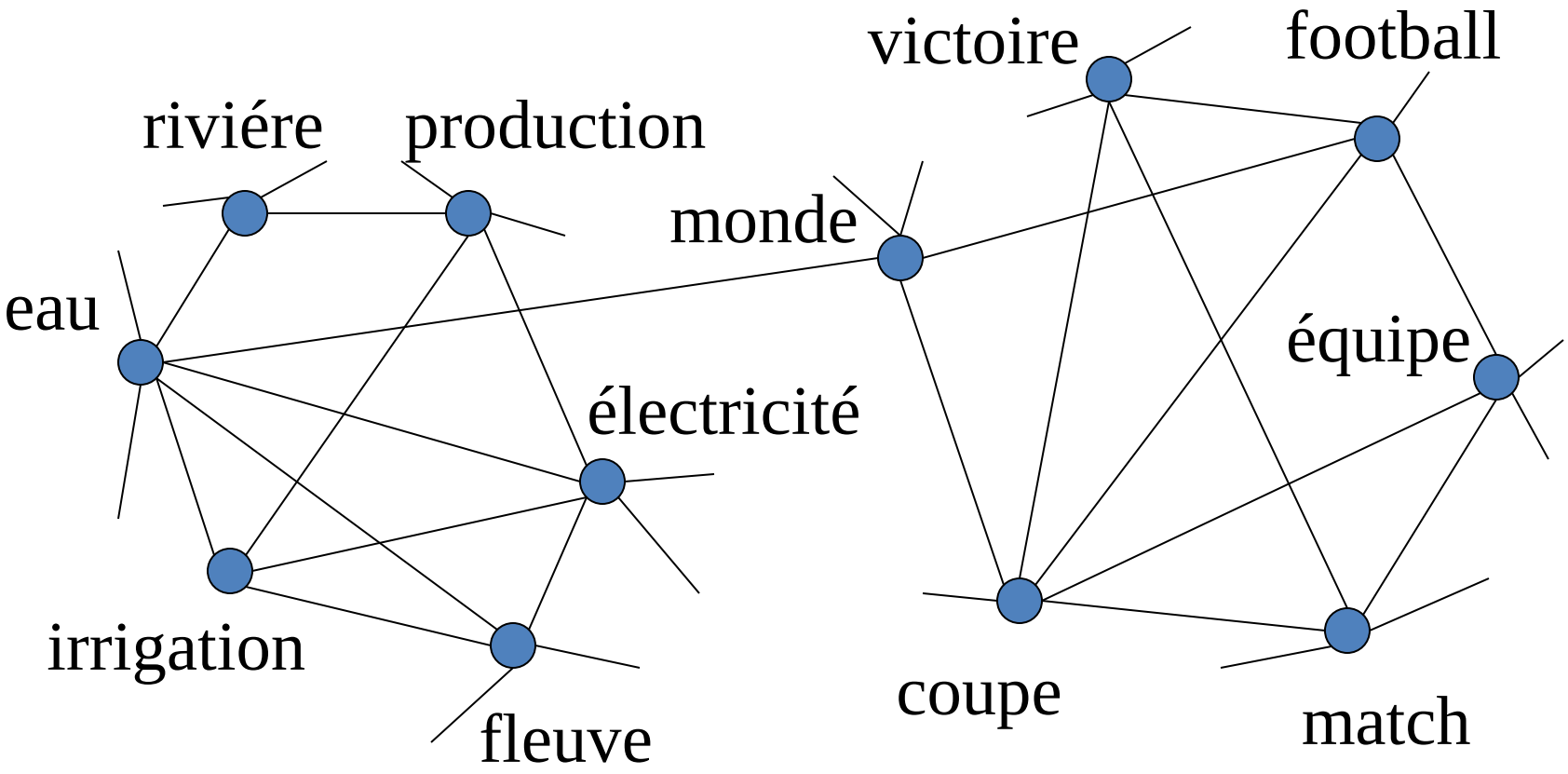
# ルートハブの見つけ方

- HyperLex
  - 高頻度で共起するノード群は同じ意味
- 処理
  1. 最も共起頻度の高いノード $n$ に対して、基準を満たしていればルートハブとする
  2. ノード $n$ に連結するノードを削除
  3. コンポーネントの作成: ルートハブと連結するノードの単語集合の作成
  4. 1.～2. を繰り返す

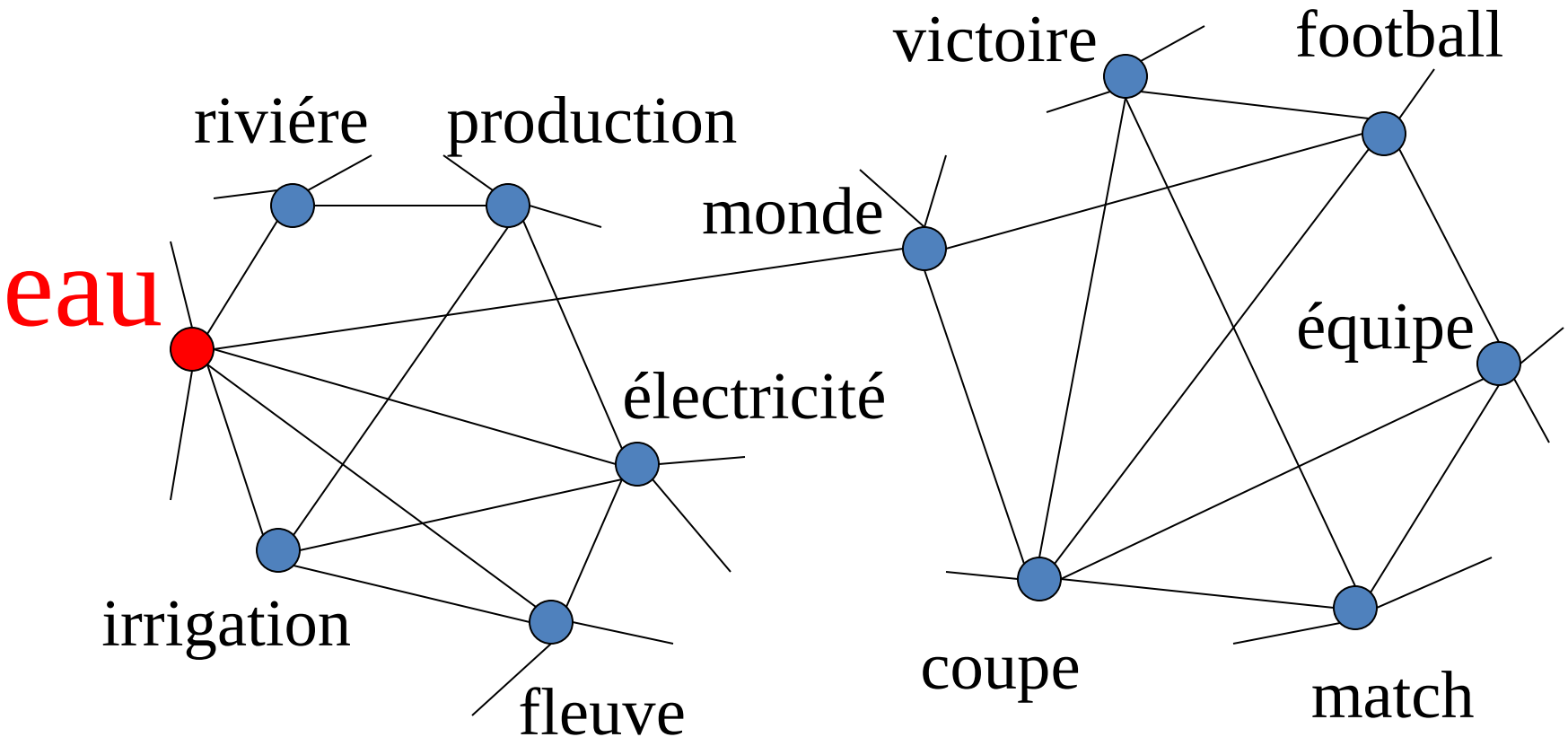
# ルートハブとする基準

1. エッジの最小頻度
2. ノードの最小頻度
3. エッジの最大重み
4. 文の最小単語数
5. ハブに隣接するノードの最小数
6. ハブに隣接するノードの平均重みの最大値
7. ハブの最小頻度
8. ハブの数

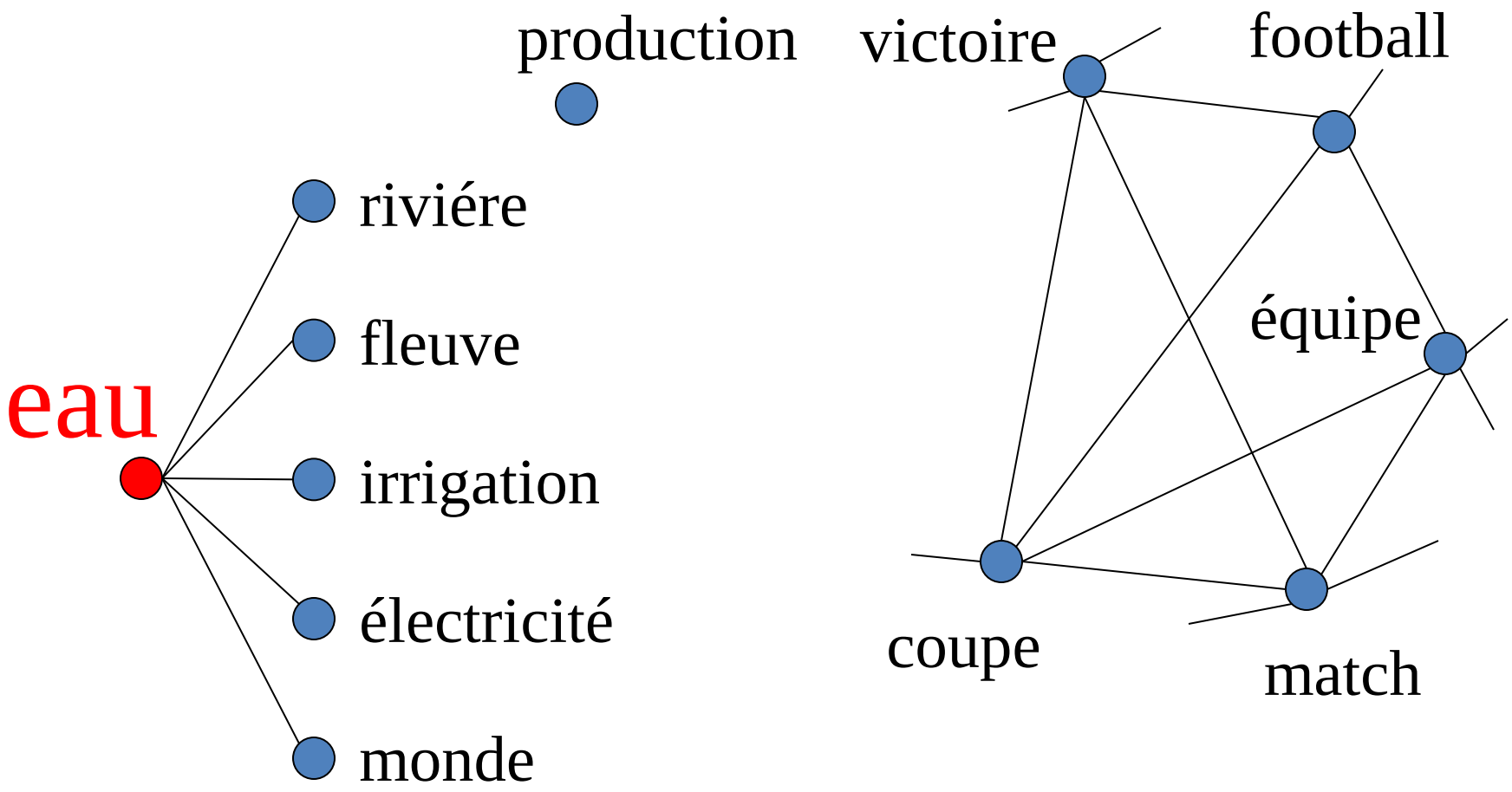
# ルートハブの例


victoire
football
riviére
production
monde
eau
équipe
électricité
irrigation
coupe
match
fleuve

# ルートハブの例

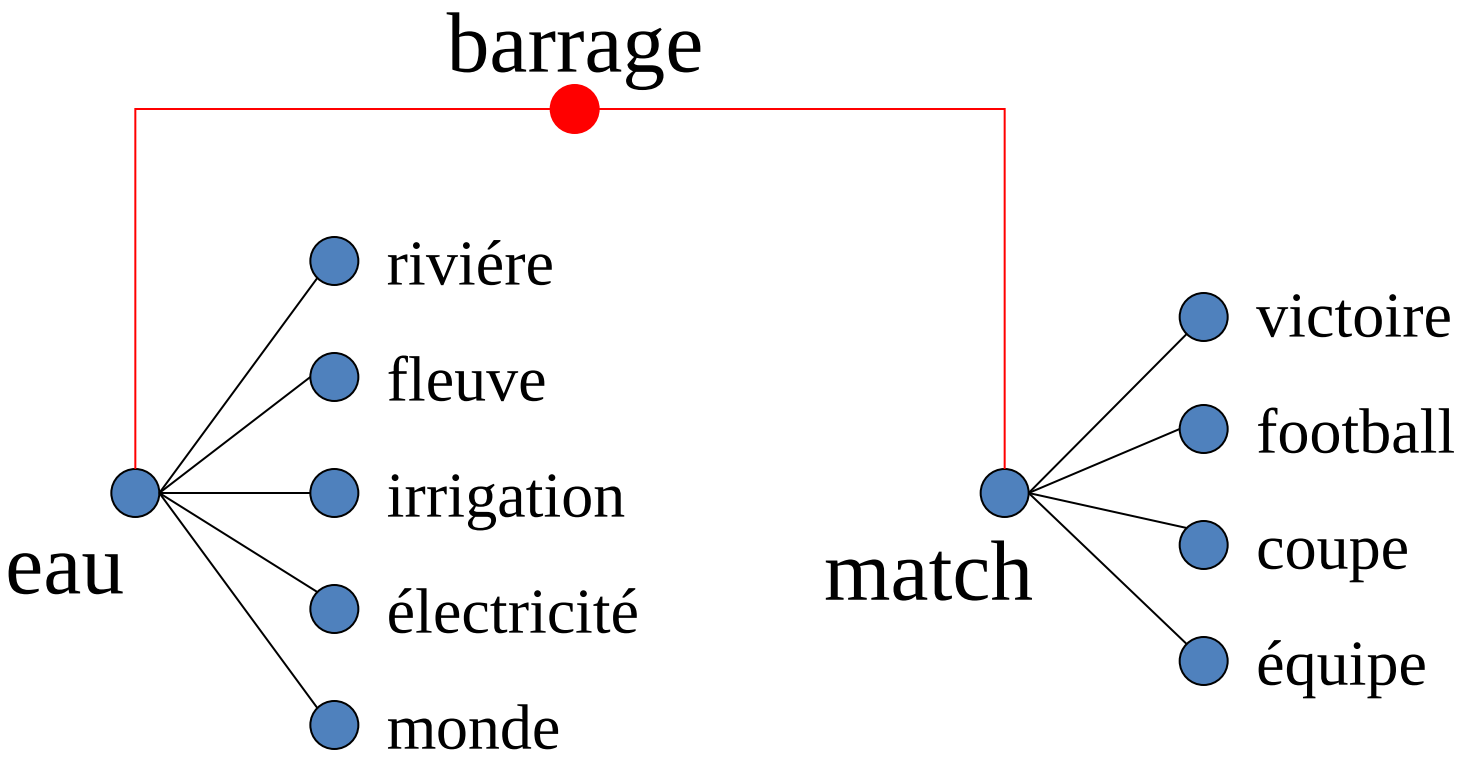

victoire
football
riviére
production
monde
eau
équipe
électricité
irrigation
coupe
match
fleuve

# ルートハブの例


production
victoire
football
riviére
fleuve
équipe
eau
irrigation
électricité
coupe
match
monde

# 最小スパン木

- 対象単語とルートハブを距離 0 で連結



# スコアベクトル

- 対象文に対し、各ハブ毎のスコアを計算
- 文中の名詞 $v$ に対してスコアを計算
  - $d(h_i, v)$: ハブ $h_i$ と $v$ の距離
  - スコア $s_i$: ハブ $i$ のスコア

$$s_i = \frac{1}{1 + d(h_i, v)}$$

- 対象文のスコアベクトル
  - すべての単語のベクトルの総和

# Markov Clustering

- 2 文の関連度を要素とする正方行列 M
  - 2 つのスコアベクトルのコサイン
  - 類似度の最大値のみを残して、その他は 0
- 行列 M に対して、マルコフクラスタリング
  - グラフ内をランダムウォーク
  - 同じクラスタならば、同じクラスタ内を巡る傾向

# おわりに

- Wikipediaによる事前知識獲得手法の提案
  - 意味ラベルとその説明文の両方を利用
  - HyperLexアルゴリズムの最小スパン木で表現
  - システムの構築と実験はこれからの課題
- 問題点
  - 固有表現には強そうだが、一般名詞は微妙
    - 「一般」、「気持ち」など項目のない単語への対応
  - 曖昧さ回避ページが存在しない単語もある
    - 「記録」、「技術」など