PHÂN LỚP VĂN BẢN TIẾNG VIỆT THEO HƯỚNG TIẾP CẬN LEXICAL CHAIN

PHẦN I:

TỔNG QUAN VỀ BÀI TOÁN PHÂN LỚP VĂN BẢN

Các phương pháp biểu diễn văn bản

- Mô hình vector
  - Văn bản = 1 vector n chiều + trọng số cho mỗi giá trị của nó
- Mô hình vector thưa
  - số từ với trọng số khác 0 nhỏ hơn rất nhiều so với số từ có trong Cơ sở dữ liệu

Các phương pháp biểu diễn văn bản

- Mô hình tần số kết hợp TF x IDF
- Xét:
  - Tập dữ liệu gồm m văn bản: D = {d_1, d_2,… d_m}.
  - Mỗi văn bản biểu diễn dưới dạng một vector gồm n thuật ngữ T = {t_1, t_2,…t_n}.
  - f_{ij} là số lần xuất hiện của thuật ngữ t_i trong văn bản d_j
  - m là số lượng văn bản
  - h_i là số văn bản mà thuật ngữ t_i xuất hiện
  - Gọi W = {w_{ij} } là ma trận trọng số, trong đó w_{ij} là giá trị trọng số của thuật ngữ t_i trong văn bản d_j

Các phương pháp biểu diễn văn bản

- Ma trận trọng số TFxIDF được tính như sau:

$$w_{ij} = \begin{cases} [1+\log(f_{ij})]\log\left(\frac{m}{h_i}\right) & \text{nếu } h_{ij} \geq 1 \\ 0 \text{ nếu ngược lại} \end{cases}$$

Các phương pháp biểu diễn văn bản (tt)

- Mô hình Lexical Chain:
  - "Lexical Chain" là một khái niệm nhằm duy trì tính cố kết giữa các từ trong văn bản có mối liên quan với nhau về mặt ngữ nghĩa
  - Một số loại quan hệ về ngữ nghĩa giữa các từ:
    - Lặp lại (Repeatation)
    - Đồng nghĩa (synonyms)
    - Trái nghĩa ()
    - Bộ phận-Toàn thể (hypernyms, hyponyms)
    - …
  - Ví dụ : ***C1= {kinh tế, thương mại, lĩnh vực, vốn, thị trường}***

Các thuật toán giải quyết bài toán Phân lớp văn bản

- Thuật toán cây quyết định.
- Thuật toán k-NN.
- Thuật toán Lexical Chain.

Thuật toán Cây quyết định

- Cây quyết định gồm các nút quyết định, các nhánh và lá :
  - Mỗi *lá* gắn với một nhãn lớp,
  - Mỗi *nút* quyết định mô tả một phép thử X nào đó,
  - Mỗi *nhánh* của nút này tương ứng với một khả năng của X.
- Ý tưởng: Phân lớp một tài liệu *dj* bằng phép thử đệ quy các trọng số mà các khái niệm được gán nhãn cho các nút trong của cây với vector cho đến khi đạt tới một nút lá => nhãn của nút lá này được gán cho tài liệu *dj*.
- Ưu điểm: chuyển dễ dàng sang dạng cơ sở tri thức là các luật *Nếu - Thì* .
- Nhược điểm:
  - Cây thu được thường rất phức tạp, chỉ phù hợp với tập mẫu ban đầu.
  - Khi áp dụng cây với các dữ liệu mới sẽ gây ra sai số lớn.

Thuật toán kNN (*K-Nearest Neighbor*)

- Tư tưởng : tính toán độ phù hợp của văn bản đang xét với từng lớp (nhóm) dựa trên *k* văn bản mẫu có độ tương tự gần nhất.
- Có 3 cách gán nhãn:
  - ***Gán nhãn văn bản gần nhất***:
  - ***Gán nhãn theo số đông***
  - ***Gán nhãn theo độ phù hợp chủ đề***
- Cách biểu diễn văn bản (hướng tiếp cận truyền thống): TF x IDF

Thuật toán Lexical Chain

- Bước 1: Đọc từ w trong văn bản.
- Bước 2: Tiến hành dừng nếu w là stop-word.
- Bước 3: Thông qua WordNet, lấy về tập S gồm tất cả các nghĩa mà w có thể có.
- Bước 4: Tiến hành tìm kiếm mối liên hệ gần nhất giữa w với các từ trong tập hợp chain đã được khởi tạo
  - Nếu tìm thấy mối liên hệ đủ gần, tiến hành kết nạp w vào chain đó, đồng thời khử nhập nhằng nghĩa cho w bằng cách tỉa đi tất cả các sense đã không được sử dụng để tìm mối liên hệ này
  - Nếu không tìm được chain nào thoả mãn, tiến hành lập chain mới và kết nạp w là từ đầu tiên.

Lý do lựa chọn hướng Lexical Chain

- Can thiệp vào bản chất ngôn ngữ của văn bản, thay vì mô hình toán học thuần tuý
- Khử nhập nhằng ngữ nghĩa của từ rất tốt.
- Hiệu quả khi hệ thống cần "học lại"
- Giúp thu gọn không gian bài toán
- Là hướng tiếp cận mới

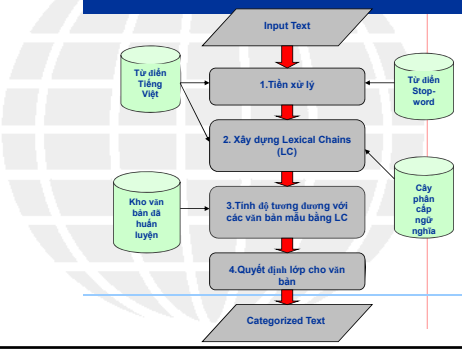
PHẦN II:

TIẾP CẬN BÀI TOÁN PHÂN LỚP VĂN BẢN TIẾNG VIỆT THEO HƯỚNG LEXICAL CHAIN

Các tác động của đặc trưng ngôn ngữ Tiếng Việt đến bài toán

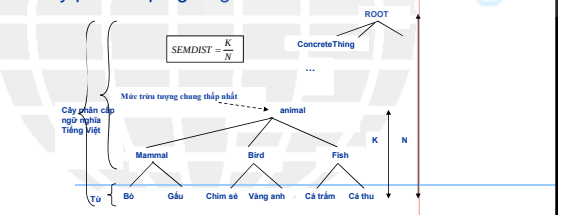
- Cần phải thiết kế thêm giải thuật để tách từ
- Không cần phải giải quyết bài toán Stemming
- Hiện tượng từ đồng âm: nhập nhằng ngữ nghĩa
- Tiếng Việt chưa có một WordNet hoàn chỉnh để biểu đạt các mối quan hệ ngữ nghĩa một cách phong phú và đầy đủ như Tiếng Anh

Mô hình giải quyết bài toán



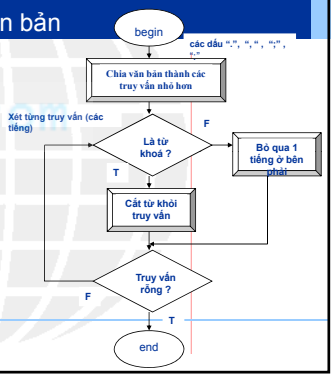
Các yếu tố ngôn ngữ được sử dụng

- Từ điển Tiếng Việt : 70.000 từ (có gắn nghĩa)
- Từ điển từ dừng
- Cây phân cấp ngữ nghĩa



Tiền xử lý văn bản

- Tách từ
- Gán nhãn từ loại, lọc ra các danh từ
- Loại bỏ từ dừng.



Giải thuật xây dựng Lexical Chain

- **Bước 1:** Với mỗi danh từ trong văn bản, liệt kê tất cả các nghĩa mà nó có thể có.
- **Bước 2:** Sử dụng WSDG để xác định nghĩa phù hợp nhất của mỗi từ trong số tập hợp nghĩa xác định ở bước 1.
- **Bước 3:** Xây dựng các Lexical Chain dựa vào nghĩa duy nhất vừa tìm được cho mỗi từ.
  - Xuất phát từ tập chain rỗng.
  - Với mỗi từ w:
    - kết nạp nó vào chain c nếu độ tương đồng của nó với tất cả các từ trong c đều đủ gần (vượt ngưỡng α lập trước)
    - Ngược lại, lập chain mới và kết nạp nó là từ đầu tiên

Đồ thị khử nhập nhằng nghĩa

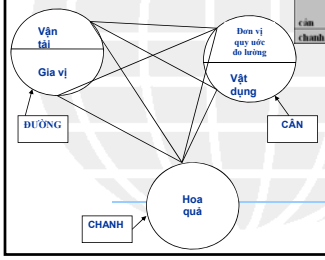
- Gọi:
  - T = {T1 , T2,… Tn} là tập các danh từ trong văn bản.
  - Si (i=1,...mi) là tập hợp các nghĩa mà danh từ Ti có thể có được (mi là số lượng nghĩa của Ti)
- G=(V,E)
  - Vi biểu diễn Ti, nhưng chia làm mi phần
  - Mỗi phần Vij biểu diễn nghĩa Sij của Ti
  - Mỗi cạnh trong E nối Vij và Vi'j'
- Mỗi cạnh được gán trọng số: $w(V_{ij}, V_{i'j'}) = sim(S_{ij}, S_{i'j'})$
- Trọng số của mỗi nghĩa Vij:

$$w(V_{ij}) = \sum w(V_{ij}, V_{i'j'}) \; (i' \neq i, i, i' = \overline{1,n})$$

Ví dụ minh hoạ giải thuật

« Sáng nay, mẹ tôi đi chợ mua hai cân đường để vắt nước chanh »

		đường		cân		chanh
		Vận tải	Gia vị	Đ vị đo lường	Vật dụng	Hoa quả
đường	Vận tải	x	x	0.3	0.3	0.2
	Gia vị	x	x	0.8	0.7	0.5
cân	Đ vị đo lường	0.3	0.8	x	x	0.7
	Vật dụng	0.6	0.3	x	x	0.5
chanh	Hoa quả	0.2	0.5	0.7	0.5	x



+ Đường: W('Gia vị') =2.0, W('vận tải') =0.8

=> Đường = Gia vị

+ Cân: W('đơn vị đo lường') =1.8, W('Vật dụng') =1.4

⇒Cân = đơn vị đo lường

Đánh giá các Lexical Chain

- Điểm cho mỗi chain:
  - score(C) = Length * Homogeneity
- Trong đó:
  - Length: Số lượng các "lượt từ" trong C.
  - Homogeneity: Tính đồng nhất giữa các từ trong C

$$\text{Homogeneity} = 1 - \alpha \frac{Number\_of\_distinct\_words\_in\_C}{Length}$$

  - Alpha = 0.75

Dùng LC tính độ tương tự giữa các văn bản

- Ký hiệu các chuỗi từ vựng c và d lần lượt là :
- *c = {c1,c2,…, cm}* và d = *{d1,d2,…, dn}*
- *Trong đó, mỗi thành phần ci, dj (i=1..m, j=1..n) đều chỉ có 1 nghĩa duy nhất lần lượt là* s_{c_i} và s_{d_j}.
- Độ tương đồng giữa c và d :

$$sim(c,d) = \sum_{i=1}^{m}\sum_{j=1}^{n} sim(s_{c_i}, s_{d_j})$$

- Độ tương tự giữa chain c và văn bản D

$$sim(c,D) = \sum_{d \in D} sim(c,d)$$

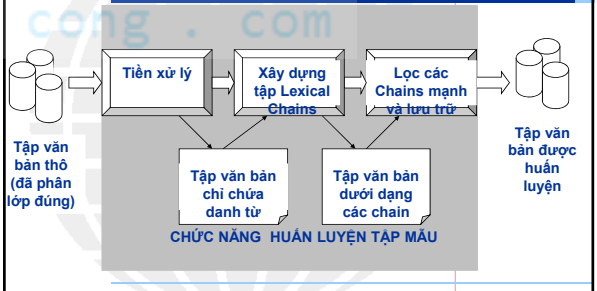
Gán nhãn lớp cho văn bản

- *Gán nhãn theo tổng độ phù hợp chủ đề*
  - Lần lượt tính tổng độ phù hợp của văn bản Q với tất cả các phân lớp có trong ***k*** văn bản đã lấy ra
  - Gán nhãn chủ đề phù hợp nhất cho Q
  - Q sẽ thuộc vào phân lớp có tổng độ liên quan cao nhất.

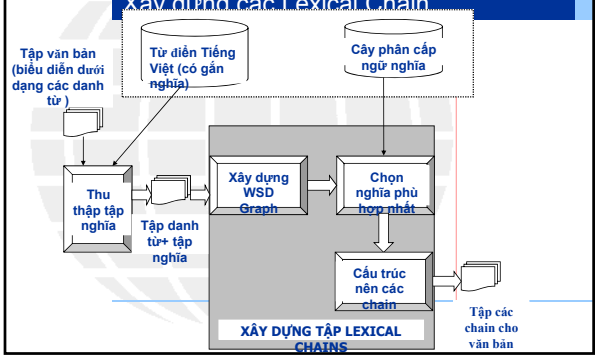
PHẦN III:

TIẾP CẬN BÀI TOÁN PHÂN LỚP VĂN BẢN TIẾNG VIỆT THEO HƯỚNG LEXICAL CHAIN

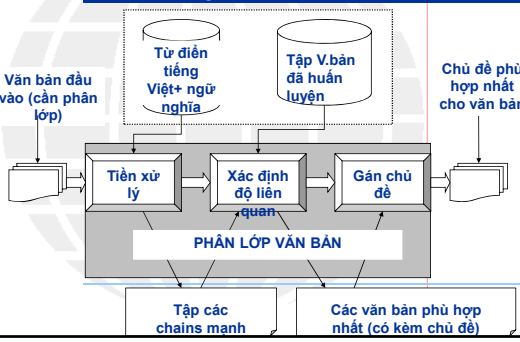
Chức năng Huấn luyện tập mẫu



Xây dựng các Lexical Chain



Chức năng Phân lớp văn bản



Thiết kế dữ liệu

➢Từ điển Tiếng Việt (nguồn: trung tâm từ điển học Vietlex):

```
<LexicalEntry>
    <HeadWord>cá quả</HeadWord>
    <Morphology>
        <WordType>composite word</WordType>
    </Morphology>
    <Semantic>
        <LogicalConstraint>
            <CategoryMeaning>Animal</CategoryMeaning>
            <Synonym>_</Synonym>
            <Antonym>_</Antonym>
        </LogicalConstraint>
        <Definition>cá dữ ở nước ngọt, thân tròn, dài, có nhiều đốm đen, đầu nhọn, khoẻ, bơi nhanh</Definition>
    </Semantic>
</LexicalEntry>
```

Thiết kế dữ liệu

➢Cây phân cấp nghĩa (nguồn: trung tâm từ điển học Vietlex):

Tên nghĩa	Dấu Tab	Root	/	Nghĩa 1	/	…	/	Nghĩa n	/	Tên nghĩa

Organization Root/ConcreteThing/LivingThing/People/Organization

Thiết kế dữ liệu

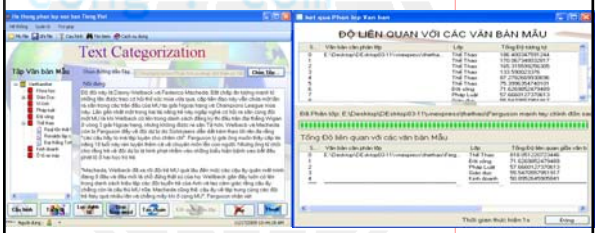
- Lưu các Lexical Chain:
  - Tập lexical chain của mỗi văn bản lưu trong một file .txt
  - Các lexical chain cách nhau 1 dòng trống
  - Trong 1 lexical chain:
    - Mỗi từ được lưu trên 1 dòng
    - Câu trúc mỗi từ như sau:

Từ	Nghĩa	Số lần xuất hiện

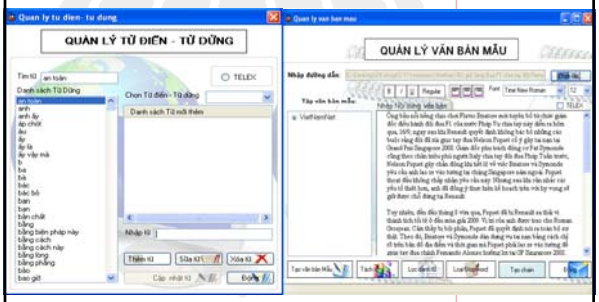
- Ví dụ:

luật sư|People|4
bị cáo|People|1
thẩm phán|People|3
cán bộ|People|2
người làm|People|1

Giao diện chính



Chức năng quản lý từ điển, từ dừng và văn bản mẫu



Tập ngữ liệu thử nghiệm

- Các bài báo được sưu tầm trên trang tin vietnamnet (http://www.vnn.vn)
- 8 chủ đề: Khoa học, Vi tính, Giáo dục, Pháp luật, Đời sống, Thể thao, Kinh doanh, Ô tô xe máy

Số bài báo	100
Số chủ đề (lớp)	8
Kích thước bài báo lớn nhất	6.13 KB
Kích thước bài báo nhỏ nhất	1.11 KB
Kích thước trung bình của một bài báo	3.30 (KB)
Số danh từ nhiều nhất trong một bài báo	89
Số danh từ ít nhất trong một bài báo	18
Số danh từ trung bình trong một bài báo	35.47

Một số kết quả thử nghiệm

Số bài báo được thử nghiệm	100
Thời gian phân lớp nhanh nhất	0.2 s
Thời gian phân lớp chậm nhất	1.9 s
Thời gian phân lớp trung bình	0.713
Số văn bản được phân lớp đúng	92
Hiệu suất phân lớp	92 %
Kích thước trung bình của mỗi bài báo	3.30 (KB)
Số danh từ trung bình trên mỗi bài báo	35.47
Số văn bản phân lớp được	100
Độ chính xác (precision)	92 %

Nhận xét

- Các văn bản bị phân lớp sai do một số nguyên nhân:
  - Bản thân nội dung văn bản cũng có sự nhập nhằng.
  - Sai từ khâu tách từ và lọc danh từ.
  - Cây phân cấp ngữ nghĩa còn hạn chế về số lượng nghĩa, dẫn đến một số danh từ có nghĩa xa nhau nhưng lại cùng thuộc về một lớp nghĩa trừu tượng (ví dụ: Concept, ConcreteThing….)
  - Độ sâu của cây chưa lớn nên dẫn tới độ tương đồng của các từ thuộc dạng trên lại cao.